Identifier les variations conduisant au cancer dans le génome non codant et du transcriptome

Jia Li

UNIVERSITE PARIS SACLAY (Paris 11)

ECOLE DOCTORALE : (Structure et Dynamique des Systèmes Vivants)

DOCTORAT

Bioinformatique

Thèse soutenue pour l'obtention

du diplôme de doctorat par

Jia LI

Identifier les variations conduisant au cancer dans le génome et transcriptome non codant

Thèse dirigée par : Professor. Daniel GAUTHERET

Soutenue le Lundi 14 Décembre 2015

JURY

Dr. Salvatore Spicuglia,	Rapporteur
Dr. Andrei Zinovyev,	Rapporteur
Dr. Hugues Roest Crollius,	Examinateur
Pr. Olivier Lespinet,	Examinateur
Pr. Daniel Gautheret,	Directeur de thèse

2015SACLS161

Acknowledgments

First of all, I would like to thank Pr**. Daniel GAUTHERET,** he accepted and supported my PhD application four years ago. Thanks for his contribution to our study, we have faced countless problems and difficulties when carrying out this work, his competent supervision and enthusiasm made our scientific hypothesis become achievable. Without his guidance and instruction, my thesis could not reach its present form.

I would like to express my heartfelt gratitude to my previous colleagues **Zohra SACI, and Cecile PEREIRA.** They gave me great help on Perl, R programming studies. Especially, Zohra SACI assisted me in learning linux command lines, perl programming and using various packages for RNA-seq analysis. Her selfless help made me quickly and successfully transit from a newbie on bioinformatics to a capable computational biologist, which greatly contributes to the success of my PhD study.

I also would like to express gratitude to my colleagues **Marie-Anne Poursat, Damien Drubay, Stefan Michiels.** Marie-Anne Poursat advised us how to correctly train and validate a random forest model, which is a critical part of our study. Damien Drubay and Stefan Michiels provided lots of useful suggestion to further improve the current study.

I also want to express gratitude to my colleagues **Claire TOFFANO-NIOCHE**, **Jean LEHMANN**, **Fabrice LECLERC**, **Nicolas CHEVROLLIER, Marc GABRIEL**. They gave me a lot of help and suggestion in my life, in particular, **Claire TOFFANO-NIOCHE,** she helped me a lot deal with various life-related problems, such as application of titre de séjour and buying medical insurance. I also want to thank my Chinese friend, Ji WANG, for his help during my PhD study in France.

Most importantly, thanks to the financial support of the China Scholarship Council (CSC) to complete my thesis in France.

At last, I want to give my sincere thanks to my family members, my father, mother and younger brother, for their caring, support and encouragement in the past three and half years.

Index

Abbreviations

CLL: Chronic Lymphocytic Leukemia

RF: Random Forest

VSURF: Variable Selection Using Random Forests

SNP: Single-Nucleotide Polymorphism

SOM: Somatic Mutation

HGMD: The Human Gene Mutation Database

Clivariant: Clinical Variant

GWAS: Genome-wide association study

ENCODE: The Encyclopedia of DNA Elements

COSMIC: Catalogue of Somatic Mutations in Cancer

CDS: Coding DNA Sequence

UTR: Untranslated Region

ncExon: Non Coding Exon

CR: Conserved Region

cTFBS: Conserved Transcription Factor Binding Site

TFBS: Transcription Factor Binding Site

DNase: DNase I Hypersensitive Site

ECS: Evolutionary Conserved RNA Structure

RR: Recombination Rate

HE: High Expression

LE: Low Expression

ER: Early Replication

LR: Late Replication

DNA met: DNA Methylation

LncRNA: Long non coding RNA

ncRNA: Non coding RNA

miRNA: microRNA

PC gene: Protein Coding Gene

RPKM: Reads Per Kilobase per Million mapped reads

FPKM: Fragments Per Kilobase Of Exon Per Million Fragments Mapped

FDR: False Discovery Rate

FI: Function Impact

RMG: Recurrently Mutated Gene

CADD: Combined Annotation Dependent Depletion

GWAVA: Genome Wide Annotation of VAriants

Chapter 1- Introduction

Results presented here are published in Cancer Letters (Appendix 1)

Mining the coding and non-coding genome for cancer drivers

.

LI J, Drubay D**,** Michiels S, Gautheret D**.** *Cancer Lett***. 2015. 369(2):307-15.**

Author contribution:

Jia LI firstly wrote the manuscript and professor Daniel Gautheret further revised the paper. Damien Drubay and Michiel Stephan gave their suggestion and comments to the work. At last, Jia LI was in charge of the final preparation and revision of this paper.

Cancer is caused by the accumulation of genetic alterations and consequent disruption of cell functions. Over the past decade, the introduction of fast and relatively inexpensive sequencing methods has provided unprecedented opportunity to characterize cancer genomic landscapes. A variety of bioinformatics tools are now available to discover genetic variations from high throughput sequencing of tumor DNA, such as GATK (DePristo et al., 2011), CRISP(Bansal, 2010), LoFreq (Wilm et al., 2012), VarScan 2 (Koboldt et al., 2012), and SNVer (Z. Wei et al., 2011), which have been recently evaluated (Pabinger et al., 2014) and (H. W. Huang et al., 2015). Depending on cancer type, tumors harbor hundreds to tens of thousands of somatic mutations, most of which are located in the non-coding portion of the genome (Lawrence et al., 2013). However, not all somatic mutations have their contributions to cancer development, they are generally divided into two main classes: the 'driver' and 'passenger' mutations. The former is causally involved in the carcinogenesis, in which it confers selective growth advantage to cancer cells and is under positive selection in the cancer microenvironment. The latter is the somatic mutation which couldn't give growth superiority to cancer cells and hasn't been positively selected, therefore, it plays little role in cancer formation and progression. Driver mutation might not be necessary for the maintenance of the final cancer but has to be selected during the cancer-evolving process. Cells which carry driver mutations and functionally inert passenger mutations undergo clonal expansion, eventually, forming the final cancer (Stratton et al., 2009). Cancer driver genes are genes which carry these driver mutations and are critical to cancer formation. They are classified into three main categories: (1) genes whose non-synonymous mutation rate is significantly greater than a background mutation rate(Lawrence et al., 2013); (2) genes accumulate mutations with high functional impact (FM bias) (Gonzalez-Perez and Lopez-Bigas, 2012); (3) genes display a higher rate of high-scoring non-synonymous mutations than silent and intronic mutations(Hodis et al., 2012) .

A critical challenge in cancer genomics study is to distinguish "driver" mutations and cancer genes that are responsible for cancer development upon specific alterations from "passenger" mutations that are mere results of the cancerous process. A number of reviews provide guidelines for the discovery of cancer-causing variants (MacArthur et al., 2014; Moreau and Tranchevent, 2012). The most common strategy is first to prioritize non-synonymous variants

in protein-coding regions and then seek recurrently mutated genes in a cohort of cancer patients (Chapman et al., 2011; Ding et al., 2008; Gui et al., 2011; Wang et al., 2011; X. Wei et al., 2011). Diverse computational methods have been explored to prioritize non-synonymous variants with respect to their disease-causing potential. Most are based on the assumption that coding mutations impacting functionally important residues, as inferred from evolutionary conservation and protein domain analysis, are more likely damaging (Vitkup et al., 2003). Other software, used in conjunction with these scoring systems, performs recurrence search in patient cohorts. Currently, 547 cancer genes are described in the COSMIC catalogue of somatic mutations in cancer (version 71) (Forbes et al., 2011a).

The immense majority of the human genome (98%) is non-coding, and consequently most somatic mutations/alterations observed in tumors occur in this non-coding fraction. Because non-coding mutations are more difficult to interpret, these regions have been mostly discounted from the wider search for driver mutations. However, mutations in non-coding regions can have a profound impact on cell fate. Indeed, functional regions in the non-coding genome include mRNA splice sites, UTR regulation elements, promoters, transcription factor binding sites, enhancers and a wide variety of non-coding RNA (ncRNA) genes. Among ncRNA genes, one particular class is now receiving focused attention due to its vast extent: long non-coding RNA (lncRNA). According to the latest estimate (Iyer et al., 2015), over 58,000 lncRNA genes are expressed in the human genome, which makes this class the biggest contributor to the "black matter" transcriptome.

There is ample evidence for disease-related mutations in the non-coding genome. A large fraction of disease or trait-relevant single nucleotide polymorphisms (SNPs) detected by Genome-wide Association Studies (GWAS) (Beck et al., 2014) is located in the non-coding genome, preferentially within enhancers, exons and mRNA promoters (Andersson et al., 2014). Inherited disease-causing variants are strongly enriched in non-coding regions under strong purifying selection, which comprise binding sites of transcription factors (TFs) and critical motifs from TF Families (Khurana et al., 2013). Further studies have shown that altered ncRNA functions initiated by genetic or regulatory changes play an important role in tumorigenesis (Chaluvally-Raghavan et al., 2014; Kwanhian et al., 2012; Ling et al., 2013; Ren et al., 2012; Tseng et al., 2014; Wegert et al., 2015).

In the absence of a clear and uniform functional code for these highly diverse non-coding elements, their variations are much more difficult to interpret than those of amino acid-coding regions. In this review we describe the methods and data available to interpret and prioritize non-coding genome mutations. As many basic principles in this field were laid for protein-coding sequence analysis, we start by reviewing the methods developed for scoring protein-coding variants. We then describe the specific non-coding elements that may be the subject cancer-driving mutations and we address the specific methods that were set up to characterize these variations.

1.1 Prioritizing coding variants

Prioritization of non-synonymous mutations for cancer study is a mature field built upon decades of experience in protein sequence and cancer pathway analysis. Table 1 provides a listing of the most commonly used tools. We distinguish below three classes of scoring systems, using either probabilistic, machine learning or hybrid approaches.

1.1.1 Probabilistic models

The pioneering SIFT (Sorting Intolerant From Tolerant) uses sequence homology to predict whether an amino acid substitution will affect protein function and hence, potentially alter phenotype (Ng and Henikoff, 2003). SIFT identifies conserved protein residues based on multiple sequence alignments of homologous proteins and calculates the likelihood that an amino acid at a position is tolerated, conditional on the most frequent amino acid being tolerated. Mutations in higher conserved coding regions intend to be predicted as more likely deleterious than those in lower conserved protein regions.

The mCluster method (Yue et al., 2010) aggregates mutation data by mapping known disease-related mutations to positions along conserved domains, and then mapping novel variants to those same conserved domains. The program identifies conserved mutation-enriched clusters, which are hotspots for cancer driving functional alterations, across multiple proteins. The mCluster score is the likelihood of a cluster of certain size occurring, given the number of positions in the domain and the mutation frequency.

MutationAssessor (Reva et al., 2011) implements a more elaborate conservation-based approach. It computes residue distribution entropy in multiple sequence alignments and estimates mutation impact by measuring the entropy difference caused by the mutation (conservation score). Moreover, the algorithm classifies protein alignment into distinct subfamilies with a clustering algorithm and quantifies the entropy difference initiated by a mutation in protein subfamilies (specificity score). The final "functional impact score" combines these two independent scores.

1.1.2 Machine learning models

PolyPhen2 (Adzhubei et al., 2010) integrates eight sequence and three structure-based attributes for the description of an amino acid substitution, and predicts the damaging effect of a coding mutation. Most PolyPhen2 features compare a property of the wild-type allele (ancestral, normal) and the corresponding property of the mutant allele (derived, disease-causing) and characterizes how likely the two human alleles are to occupy the site given the pattern of amino-acid replacements in a multiple-sequence alignment. The probability of a deleterious allele replacement is predicted using a Naïve Bayes classifier trained on HumDiv and HumVar (Capriotti et al., 2006), two databases of damaging alleles.

CHASM uses a random forest classifier to discriminate driver missense mutations from synthetically generated passenger mutations (Carter et al., 2009). It includes 49 predictive features ranging from exon conservation to UniProt annotation and frequency of the missense change type in the COSMIC database of cancer mutations (Forbes et al., 2011a). The program computes a classification score for each missense mutation. A mutation is determined to be driver or passenger by comparing its score to a null distribution made of scores from a filtered set of synthetic passengers that were held out from the Random Forest training.

SNAP (Screening for Non-acceptable Polymorphisms) is a neural network-based tool that predicts the effect of a missense variant (Bromberg and Rost, 2007). It uses PMD (the Protein Mutant Database) (Sjöblom et al., 2006) and incorporates evolutionary constraints, transition frequencies for mutations, biophysical characteristics of the substitution, secondary structural

information, relative solvent accessibility, and SwissProt annotations information to build a neural network model, which is trained on known mutations from PMD.

MutPred (Li et al., 2009) is another Random Forest classifier trained on five databases of human amino acid substitutions, CANCER (Sjöblom et al., 2006), KINASE (Greenman et al., 2007), The Human Gene Mutation Database (HGMD)(Stenson et al., 2009), Swiss-Prot (Boeckmann et al., 2003) and a broad array of attributes describing structure features (such as secondary structure, solvent accessibility), a variety of functional sites (such as DNA-binding or phosphorylation sites), evolutionary conservation and transition frequencies. The MutPred model then associates a given non-synonymous mutation to a probability of gain or loss of structural and functional features.

1.1.3 Hybrid models

The current trend for increasing the accuracy of impact measure is to integrate different methods. For example, CanPredict (Kaminker et al., 2007a) uses a random forest classifier to predict whether a change is likely to be cancer-associated, based on analyses of three scores: the SIFT score determining functional impact of change, the Pfam-based LogR.E-value metric (Clifford et al., 2004) and the Gene Ontology Similarity Score (GOSS), which measures how similar a given mutated gene is to known cancer-causing genes (Kaminker et al., 2007b).

Condel (González-Pérez and López-Bigas, 2011) combines the output from PolyPhen2, SIFT, Mutation Assessor, Pfam-based LogR.E-values and MAPP (Stone and Sidow, 2005), which predicts deleterious mutations based on their disruption of physicochemical protein characteristics. Another hybrid tool, CoVEC (Consensus Variant Effect Classification) (Frousios et al., 2013) integrates prediction results from SIFT, PolyPhen2, Mutation Assessor and SNPs&GO (Calabrese et al., 2009), a scoring system based on functional protein features such as sequence conservation and GO-terms. Finally, Combined Annotation scoRing toOL (CAROL) combines the scores of PolyPhen-2 and SIFT to predict the effect of non-synonymous coding variants (Lopes et al., 2012). Expectedly, the authors of Condel, CoVEC and CAROL demonstrate that these tools outperform most individual

methods in classifying variants as damaging or neutral, highlighting the benefits of combined approaches (Frousios et al., 2013; González-Pérez and López-Bigas, 2011; Lopes et al., 2012).

1.1.4 Comparing coding mutation scoring tools

The authors of CoVEC (Frousios et al., 2013) assessed the classification performance of their tool and nine other prediction softwares: SIFT, PolyPhen2, SNPs&GO, PhD-SNP, PANTHER, Mutation Assessor, MutPred, Condel and CAROL. Based on the programs' ability to properly classify HGMD inherited disease-related variants (Stenson et al., 2009) and neutral SNPs, MutPred had the best performance in terms of true positive rate, followed by PolyPhen2. SNPs&GO showed most applicability in cases requiring minimal false positive rates. Most of the individual tools had similar overall (ROC curve-based) performances, however, combined tools such as CoVEC were shown to outperform the individual tools. In an independent benchmark, Thusberg et al (Thusberg et al., 2011) tested nine scoring tools for their ability to distinguish 40,000 pathogenic variants of the PhenCode database (Giardine et al., 2007) from neutral variants. Tested tools included MutPred, Panther, PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP, SNPs&GO and nsSNPAnalyzer (Bao et al., 2005). Programs SNPs&GO and MutPred had best overall prediction accuracy.

	Based on	Machine learning	Cancer-specific	Other tools used	Web server, references
SIFT	Conservation	Alignment scores	No		http://sift.jcvi.org/(Ng and Henikoff, 2003)
Polyphen 2	Conservation Structure Training set	Bayesian classification	No		http://genetics.bwh.harvard.edu/pph2/)(Adzhubei et al., 2010)
Mutation assessor	Conservation		No		http://mutationassessor.org/(Reva et al., 2011)
CHASM	Conservation Structure Annotation Training set	Random Forest	Yes		http://wiki.chasmsoftware.org/index.php/MainPage(Carter et al., 2009)
mCluster	Training set		Yes		http://www.mcluster.org(Yue et al., 2010)
SNAP	Conservation Structure annotation Training set	Neural network	No	Gene ontology	http://rostlab.org/services/snap/(Bromberg and Rost, 2007)
Canpredict	Conservation Annotation	Random forest	Yes	SIFT LogR.E GOSS	http://research-public.gene.com/Research/genentech/canpredict/ (Kaminker et al., 2007a)
MutPred	Conservation Structure Annotation Training set	Random forest	No	SIFT	http://mutpred.mutdb.org/(Li et al., 2009)
Condel	Hybrid scoring system (weighted score)	NA	No	PolyPhen2, SIFT, Mutation Assessor, Pfam-based LogR.E-values and MAPP	http://bg.upf.edu/fannsdb/(González-Pérez and López-Bigas, 2011)
CoVEC	Hybrid scoring system	SVM	No	SIFT, PolyPhen2, SNPs&GO, Mutation Assessor	http://www.dcs.kcl.ac.uk/pg/frousiok/variants/index.html (Frousios et al., 2013)
CAROL	Hybrid scoring system	No	No	SIFT, PolyPhen2	http://www.sanger.ac.uk/resources/software/carol(Lopes et al., 2012)
nsSNPAnalyzer	structural and evolutionary information	Random forest	No		http://snpanalyzer.utmem.edu/(Bao et al., 2005)
PANTHER	Conservation	Alignment scores	No		http://www.pantherdb.org/tools/csnpScoreForm.jsp(Thomas et al., 2003)
PhD-SNP	Conservation Training set	Support vector machine	No		http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi (Capriotti et al., 2006)
SNPs&GO	Conservation Swissprot features	Support vector machine	No		http://snps-and-go.biocomp.unibo.it/snps-and-go/(Calabrese et al., 2009)
MAPP	Physicochemical constraints	NA	No		http://mendel.stanford.edu/supplementarydata/stone_MAPP_2005/ (Stone and Sidow, 2005)
IntOGen-mutations	Hybrid scoring system	NA	Yes	PolyPhen2, SIFT, Mutation Assessor	http://www.intogen.org/web/mutations/v04/search (Gonzalez-Perez et al., 2013)

Table 1. Summary of computational methods for predicting the effects of missense mutations in cancer.

1.2 Integrating recurrence for driver prediction

Further to prioritizing individual mutations as shown above, a variety of approaches predict driver genes by combining mutation scores and recurrence patterns. The assumption underlying these methods is that genes critical to the development of a specific cancer type should be recurrently mutated in a cohort of cancer samples. Several programs are available to identify such genes (Chapman et al., 2011; Ding et al., 2008; Gui et al., 2011;Wang et al., 2011;X. Wei et al., 2011).

IntOGen-mutations is a web server aiming to identify cancer drivers across tumor types (Gonzalez-Perez et al., 2013). The system first determines the consequences of mutations using the Ensembl variant effect predictor tool which offers a comprehensive database of variations, their effects and context (Chen et al., 2010) and uses three of the above tools (SIFT, PolyPhen2 and MutationAssessor) to compute the functional impact score of a somatic mutation. These functional scores are then transformed into a uniform score which measures the damaging impact of somatic mutations with transFIC (González-Pérez and López-Bigas, 2011). This pipeline also computes each mutation's frequency of occurrence within and across cancer projects and groups mutations occurring in the same gene (or pathway). Subsequently, OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012) which detects genes accumulating mutations with high functional impact (FM bias) and OncodriveCLUST tools (Tamborero et al., 2013) which determine genes whose mutations cluster in particular regions of the protein sequence in comparison with synonymous mutations (CLUST bias) are used to identify positively selected genes, *i.e.* genes whose mutations are selected during tumor development and are therefore likely drivers. Finally, the pipeline computes the frequency of mutation of each gene (and pathway) within a cancer class (Figure1).

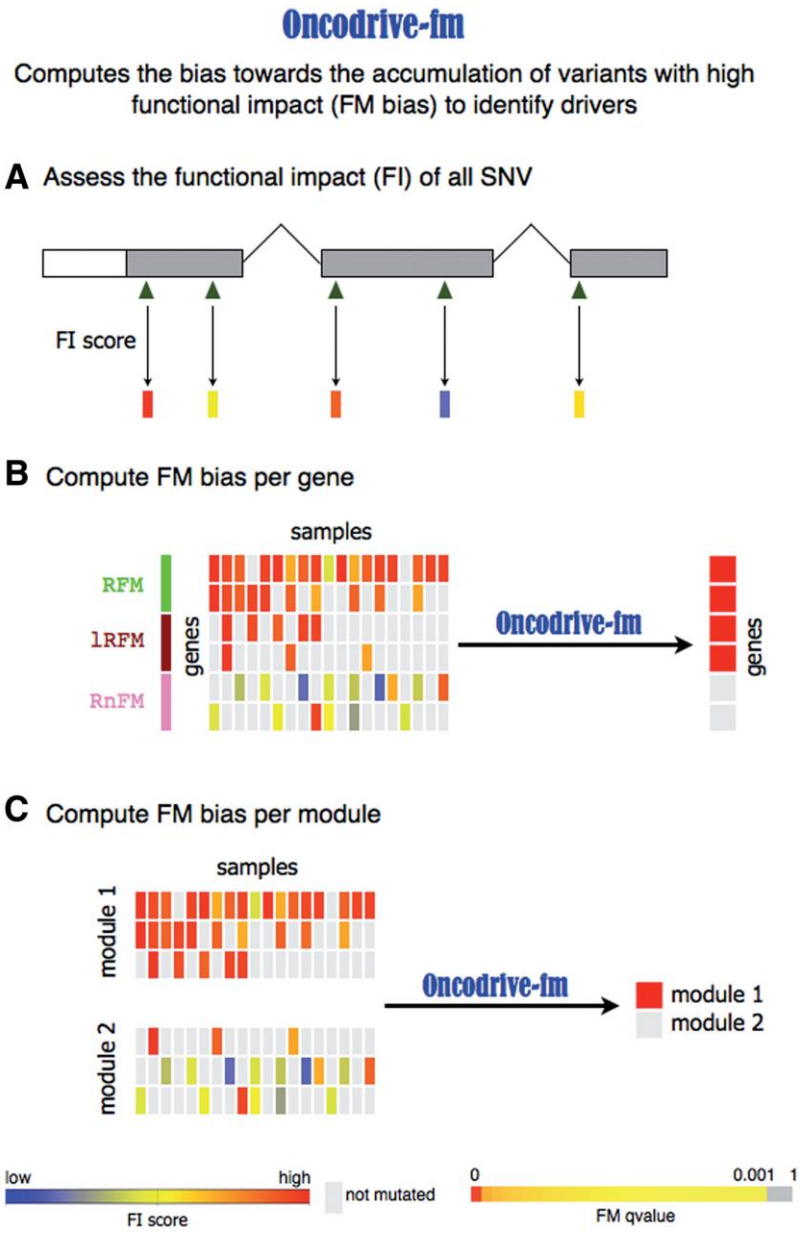

Figure1. Schematic display of the Oncodrive-fm driver detection tool (Gonzalez-Perez and Lopez-Bigas, 2012).

Oncodrive-fm is constructed based on the hypothesis that driver genes display the bias toward the enrichment of variants with high function impact (FI). (A) The first step of Oncodrive-fm is measurement of FI scores of coding variants detected in multiple cancer samples with SIFT, polyphen2 and MutationAssessor. (B) Secondly, Oncodrive-fm evaluates whether a gene possesses a shift toward the enrichment of variants with high FI, it compares the FI of observed variants to a null distribution and computes a P-value for each gene. RFM, Recurrent and FM biased; lRFM, Lowly Recurrent and FM biased; RnFM, Recurrent but not-

FM biased. (C) Lastly, Oncodrive-fm can also detect gene modules or pathways that possess the FM bias.

The MutSigCV method (Lawrence et al., 2013) assesses the background mutation rate for each gene–patient–category combination based on the observed silent mutations in the gene and non-coding mutations in the surrounding regions. It pools data from other genes with similar properties (for example replication time, expression level) to increase accuracy. Significance levels (P values) are determined by examining whether observed mutations in a gene significantly exceed the expected counts based on the background model (Figure2).

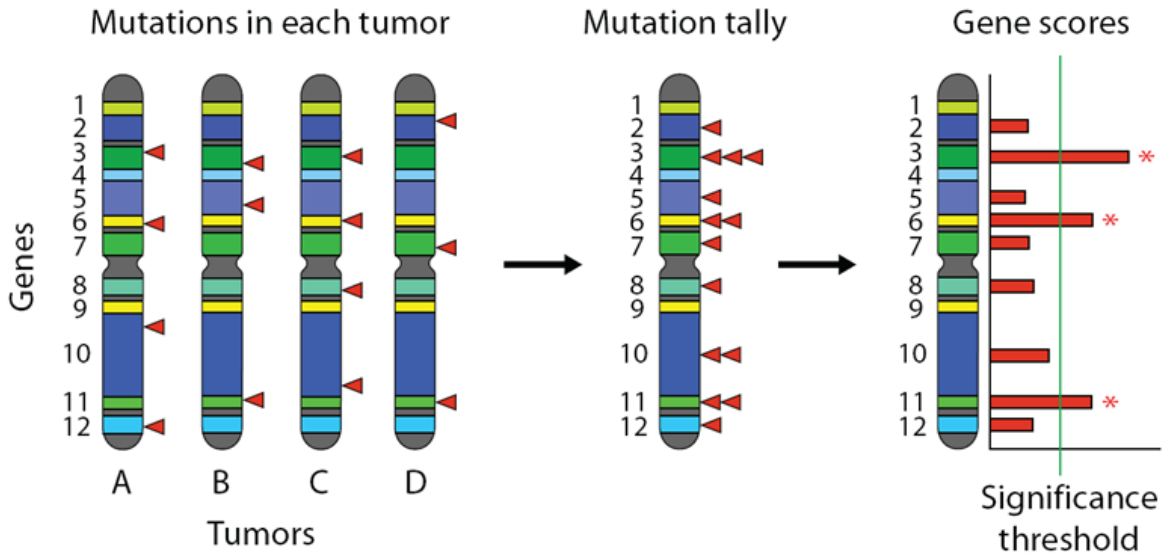

Figure2. Overall concept of detection of recurrently mutated genes of MutSigCV in a cohort of cancer samples (Lawrence et al., 2013).

MuSiC relies on the calculation of a background mutation rate (BMR) (Dees et al., 2012). The algorithm counts the number of bases with sufficient aligned read-depth based upon user-defined coverage. Counts are determined for A, T, C and G as CpG dimers, and non-CpG C and G. Discovered mutations are categorized according to mutational mechanism, with separate categories for AT transitions, AT transversions, CpG transitions, CpG transversions, CG (non-CpG) transitions and transversions, and a seventh 'indel' category. The BMR of each mutational mechanism category is calculated by dividing the number of mutations found in that category by the total number of bases available in which such a call could have been

made. Significantly mutated genes are generated by comparisons of mutation rates to BMR, using statistical tests.

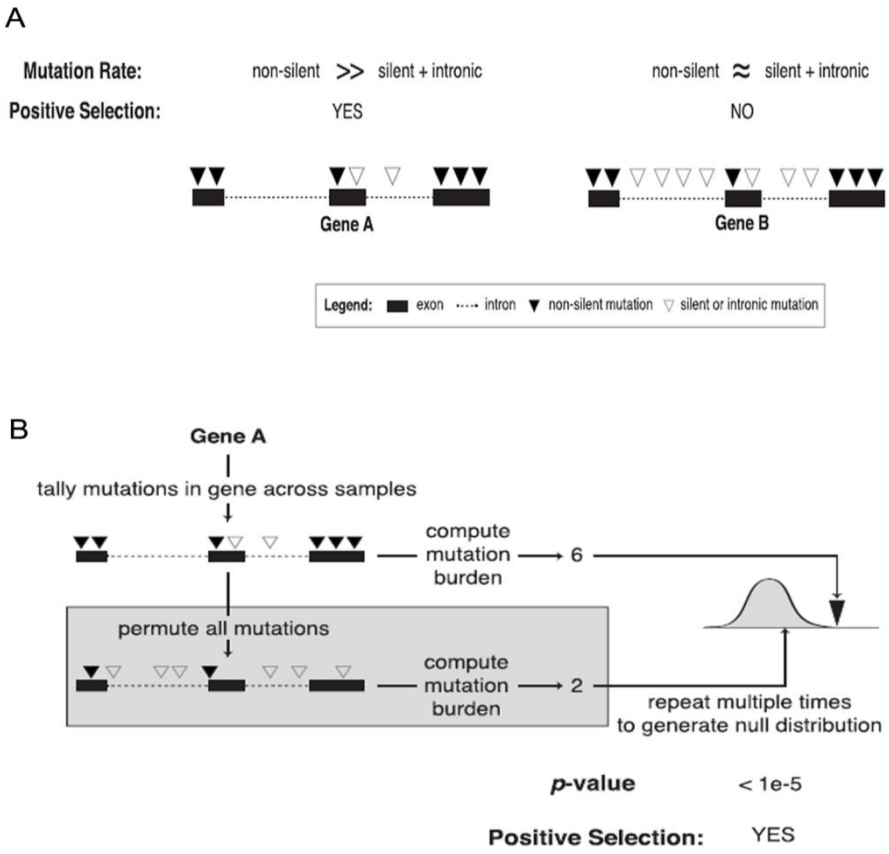

Figure 3. Identification of driver genes under positive selection with InVEx (Hodis et al., 2012) (A) Gene A possesses higher rate of nonsilent variants and silent/intronic variants in comparison with that of Gene B, indicating gene A is under positive selection of nonsilent variants in cancer. (B) Schema of a random permutation-based approach to prioritize driver genes that possess positively selected nonsilent mutations with respect to a null distribution.

InVEx is a random permutation-centered algorithm (Hodis et al., 2012) that relies on the assumption that a gene under positive selection for nonsilent mutations during cancer formation displays a higher rate of high-scoring non-synonymous mutations than silent and intronic mutations. A random permutation test is performed across each gene and a "mutation burden" score is calculated for each randomized instance, providing a null model of score

distribution. The actual mutation burden observed for a gene across all samples is then compared to this distribution and a P-value is computed, assessing whether the observed coding mutations and genes undergo positive selection (Figure3).

Although genes that are mutated with high recurrence are easily recognized, some cancer drivers are mutated in a small fraction (*e.g.* <1%) of tumors (Wood et al., 2007). Thus, methods that can classify mutations as either drivers or passengers on the basis of data that is independent of mutation frequency clearly become important. There are many ways of combining mutation deleteriousness, recurrence and knowledge of mutational background. Computational options in this area are far from fully explored and we may thus expect improved driver predictors in the future. Furthermore, the application of these methods to the non-coding genome is a fascinating perspective, as so little is known about driver elements in these regions. This challenge may soon become accessible thanks to development of scoring systems for non-coding mutations, as explained in the next sections.

1.3 Non-coding elements and cancer

The list of non-coding elements involved in gene expression regulation has been steadily increasing over the years. Promoters, enhancers, splicing regulators and the expanding family of regulatory ncRNA (mainly miRNAs and lncRNAs) are central elements of the cell regulatory network. Their function in the control of gene expression is similar to that of many protein-coding cancer drivers, half of which are involved in transcriptional and posttranscriptional regulation. Therefore, it comes as no surprise that mutations within these non-coding elements are responsible for the initiation and progression of cancer, among other diseases (Andersson et al., 2014; F. W. Huang et al., 2013; Khurana et al., 2013; Killela et al., 2013; Horn et al., 2013).

The first non-coding cancer hotspots to be suspected were promoters and TF binding sites. Indeed, among 4,492 phenotype-associated SNPs from the GWAS Central Database (Beck et al., 2014), 12% are located in binding regions of transcription factors, which is significant as these loosely defined regions represent 8.1% of the genome (Sato et al., 2013). Genetic variations at TF binding sites, including single-nucleotide polymorphisms and larger

structural variants, are frequently associated with binding affinity (Kasowski et al., 2010; Mcdaniell et al., 2010; Zheng et al., 2010), gene expression (Sugimachi et al., 2014; French and Et Al, 2013) and cancer susceptibility, progression and outcome (Jiang et al., 2012; Lin et al., 2014; S.-P. Huang et al., 2013). A well-known such locus is the TERT promoter, whose mutations were established as drivers in melanomas and gliomas (Killela et al., 2013; F. W. Huang et al., 2013; Horn et al., 2013).

Another important class of regulatory element is that of splicing regulators. Misregulation of RNA splicing initiated by genetic variants is a cause of human disease, including cancer. Alteration of 5'and 3' splicing sites and adjacent bases accounts for 10% of human inherited disease mutations (Sterne-Weiler and Sanford, 2014; Krawczak et al., 2007) and the number of tumor-relevant splicing variants detected by GWAS in cancers reaches 15,000 (He et al., 2009; Venables et al., 2008; Shapiro et al., 2011). For example, a germline mutation in the splicing site of hSNF5 is causative of exon 7 skipping and subsequent frameshift, which, as a result, renders infants susceptible to develop malignant brain tumors (Taylor et al., 2000). Likewise, a mutation at the acceptor site of the APC gene intron 3–exon 4 junction causes the loss of exon 4, which accordingly terminates seven codons downstream of junction 4, a phenomenon closely associated to childhood hepatoblastoma (Kurahashi et al., 1995).

Variation in non-coding RNA (ncRNA) sequence and expression is another potential component of cancer progression. The first important offenders in this class were miRNAs. Single nucleotide variations in miRNA sequences or in their mRNA target sites lead to alteration of binding specificity, thus affecting expression and/or translation of target mRNAs (Manikandan et al., 2012; Gopalakrishnan et al., 2014; Kamaraj et al., 2014; Manikandan and Munirajan, 2014; Vaishnavi et al., 2014). For instance, SNPs in mRNAs of the CEP family of cell division genes, alter mRNA/miRNA interactions, greatly affecting mRNA expression, disrupting the cell cycle and contributing to initiate cancer(Kamaraj et al., 2014) Overall, more than 236 miRNAs have been associated to 79 human cancers either as potential oncogenes or tumor suppressors (Xie et al., 2013).

Long non-coding RNA is the most recent class of regulatory ncRNA to be associated to cancer. According to a recent study (Iyer et al., 2015), over 68% (58,648) of expressed genes

in human tumors are lncRNAs, 7942 of them lineage- or cancer-specific. Through gene regulation or other mechanisms, lncRNAs may act as proto-oncogenes, tumor suppressor genes or drivers of metastatic transformation. For instance, the HOTAIR lncRNA is highly expressed in primary breast tumors and metastases, as well as in gastric cancer, and its repression inhibits xenograft tumor growth and metastasis in mouse models (Gupta et al., 2010; Okugawa et al., 2014). MALAT1 is another lncRNA whose expression is correlated with metastasis and survival in lung cancer (Ji et al., 2003). Knockout of MALAT1 greatly impairs the migration and formation of tumor nodules of MALAT1-deficient A549 cells in a mouse xenograft (Gutschner et al., 2013). Jin et al. (Jin et al., 2011) observed that among a set of 33 SNPs independently associated with elevated prostate cancer (PCa) risk, eight were located in lncRNAs. Moreover, lncRNA loci showed a five-fold enrichment of PCa risk-related SNPs in comparison with the entire genome. SNPs in the lncRNA PRNCR1 were proposed to be related to colorectal cancer (CRC) risk (L. Li et al., 2013).

In spite of these recent advances, the list of cancer-driving elements in the non-coding genome remains extremely short with respect to the size of the regions involved. A major avenue in identifying new potentially relevant loci involves exploring chromatin states. Indeed, regions where chromatin is open or active in a given cell type are the most likely to contain key regulatory elements. For instance, DNase I hypersensitive sites (DHSs),*i.e.* DNA regions sensitive to the DNase I enzyme, harbor many regulatory elements such as enhancers, promoters and silencers (Gross and Garrard, 1988; He et al., 2014). Moreover, DHSs are associated with elevated levels of nearby gene expression, at least in certain cells (He et al., 2014). Other important functional hallmarks are provided by histone modifications such as acetylation and methylation, which control chromatin states and are thus important regulators of gene expression (Dawson and Kouzarides, 2012). Specific histone marks suggest different types of regulatory elements: H3K4me3 generally marks promoters and transcription start sites. Putative enhancers tend to be marked with H3K4me1 alone or in combination with H3K27ac or H3K27me3 (Rada-Iglesias et al., 2011; Zentner et al., 2011). Conversely, major repressive marks, such as H3K9me3 and H3K27me3, are associated with constitutive heterochromatin and repetitive elements, repressive domains and silent developmental genes (Rada-Iglesias et al., 2011) and are therefore less likely to harbor cancer drivers.

1.4 Prioritizing non-coding variants

Although the number of cancer-associated non-coding mutations is increasing, finding cancer-driving mutations in the non-coding genome remains a huge challenge. A major bottleneck lies in identifying functional domains while trying to explore the consequences of the variations. Functional interpretation of non-coding variations is now turning into a realistic goal through the completion of major high-throughput studies such as the Encyclopedia of DNA Elements (ENCODE) (Rosenbloom et al., 2013), the "29 Mammals" Project (Lowe and Haussler, 2012), the Health Roadmap Epigenomics project (Bernstein et al., 2010) and other large scale regulatory data collections (Rhee and Pugh, 2011; Yu et al., 2011; Zeller et al., 2010)(Degner et al., 2012; Palii et al., 2011). Particularly, The ENCODE Project has provided researchers with genome-wide mapping of histone modification, Dnase I hypersensitive sites, FAIRE sites (formaldehyde-detected nucleosome-depleted elements), transcription factor binding sites, RNA-seq expression data and replication timing across multiple cell lines (Rosenbloom et al., 2013). These extensive data form a major stepping-stone toward the functional annotation of non-coding variants. More and more studies are taking advantage of these annotations to explore and prioritize non-coding variants implicated in cancer and other diseases. Table 2 presents seven systems that are currently available for scoring non-coding variants. We distinguish below two families of such methods, based either on empirical scoring systems or on machine learning.

1.4.1 Empirical scoring systems

The RegulomeDB database and software (Boyle et al., 2012) assigns functions to non-coding variants based on the principle that a variant impacting a regulatory element likely results in functional consequence. Non-coding variants are classified into different functional categories according to their overlap with functional elements such as transcription factor binding, histone modifications, DNase I hypersensitive sites, FAIRE sites and eQTLs (expression Quantitative Trait Loci, that is loci likely to affect expression of target genes). Application of this tool to the annotation of non-coding variants from 69 full sequenced genomes (Clarke et al., 2012) identified thousands of potential functional variants.

The FunSeq tool (Khurana et al., 2013) predicts non-coding drivers by scoring the deleterious potential of variants, based on two assumptions. First, somatic variants in non-coding elements containing a high fraction of rare variants (derived allele frequency < 0.5%) are considered as under negative selection and thus are most likely to be cancer drivers. Second, driver mutations should be recurrent in the same genomic element across multiple cancer samples. Application of this workflow to 90 cancer genomes yielded nearly a hundred non-coding drivers candidates. An improved algorithm, FunSeq2 (Fu et al., 2014) exploits large-scale genome data from 1000 Genomes and ENCODE into a scoring pipeline that combines functional features such as sequence conservation, transcription-factor binding sites, enhancer-gene linkages, network centrality and recurrence across samples. In this model, features are weighted by their probability of overlapping a natural polymorphism in the 1000 Genome database, which is a negative indicator of selection strength. Application of FunSeq2 to germline pathogenic regulatory variants successfully distinguished HGMD (Human Gene Mutation Database) and GWAS non-coding pathogenic variants from neutral ones. The method also effectively scored COSMIC recurrent variants higher than non-recurrent variants.

1.4.2 Machine-learning models

While the RegulomeDB and FunSeq systems prioritize functional genetic variations using empirical models, recent methods aim to integrate functionally predictive features automatically using machine learning (Kircher et al., 2014; Ritchie et al., 2014; Shihab et al., 2015). One of these models, GWAVA (Ritchie et al., 2014) uses regulatory mutations annotated in the HGMD database as a training set for non-coding variants of medical importance. These variants are predicted using a random forest classifier based on a combination of regulatory features, genic context and genome-wide properties such as DNase I hypersensitivity sites, FAIRE sites, Transcription factor binding sites, Histone modifications, RNA polymerase binding sites, complex epigenetic states, CpG islands, sequence conservation, allele frequency of variants and gene annotation. The model was able to effectively discriminate a set of disease-relevant variations of the ClinVar (Landrum et al., 2014) and GWAS Central databases from control variants. More importantly, recurrent cancer mutations from the COSMIC database were scored significantly higher than non-recurrent

mutations, suggesting that this approach might be useful in prioritizing cancer driver mutations.

Another tool, FATHMM-MKL, implements multiple kernel learning to weight different ENCODE feature annotations based on their relevance. The program builds a Support Vector Machine classifier based on a positive training set of non-coding pathogenic variants annotated in HGMD and a negative set of common single-nucleotide variants with allele frequency above 1% within 1-Kb surrounding disease-causing variants. The model uses for prediction a kernel matrix of 10 annotation features, including transcription factor binding sites, evolutionary conservation, DNase I hypersensitive sites and histone modifications (Shihab et al., 2015). A possible limitation in GWAVA and FATHMM-MKL is the methods highly rely on a set of promoter proximal, pathogenic mutations that are well characterized and thus are subject to ascertainment bias.

Instead of building a classifier using limited curated pathogenic variants, the CADD system (Kircher et al., 2014) contrasts the annotations of fixed derived alleles in humans with those of de novo simulated variants. Here fixed (or nearly fixed) alleles are used as models for deleterious variants. The CADD system is trained to recognize such variants using a support vector machine classifier based on a combination of 63 tracks of annotations, including conservation, regulatory information, transcript information, protein-level score produced by SIFT, Polyphen or Grantham (Grantham, 1974). CAAD successfully differentiated 14.7 million high-frequency human-derived alleles (observed variants) from 14.7 million simulated variants (half simulated de novo mutations).

To conclude this section, we mention SPANR (splicing-based analysis of variants) (Hs et al., 2015), a program that combines a Bayesian machine learning algorithm and a regulatory model of gene splicing to detect and score disease-associated genetic variants. The RNA splicing model integrates regulatory elements and splicing levels generated from RNA-seq data of healthy human tissues. SPANR is capable of a precise classification of both intronic disease-related variants and deleterious disease mutations within exons, from common variants in the dbSNP database. Analyses using SPANR have generated a large body of

splice-disruptive mutations involved in Autism, familial colorectal cancer and spinal muscular atrophy, which are known for RNA-splicing deregulation.

1.4.3 Comparing non-coding variant scoring tools

To illustrate the divergence of predictions by different non-coding mutation scoring systems, we selected seven tools from the current literature (CAAD, FunSeq, FunSeq2, GWAVA, RegulomeDB, Fathmm-MKL and SPANR) and used them to score 874,325 non-coding variants (both substitutions and short indels) from the whole genome sequencing of 88 liver cancer samples (Lawrence et al., 2013). First, we should note that all tools are not applicable to the entire set of somatic mutation (Fig. 4A). GWAVA, RegulomeDB, and funSeq2 were able to score over 99% of variants, while SPANR provided scores for only 2.48% of variants due to its specificity for splicing regulation. Due to this different scope, we excluded SPANR from further comparison. We scored the 841,402 somatic mutations covered by the other 5 tools and collected the 10,000 highest scoring variants from each tool. The Venn diagram in Fig. 4B shows the overlapping of predictions. Strikingly, even though there is a higher overlap of highest scoring variants among five tools as compared to 10000 randomly sampled ones (P value=0, a permutation test), only 13 variants are commonly predicted as high scoring by all five tools, illustrating the remarkable divergence of non-coding variant prioritization strategies. While a full benchmark of the different prediction algorithms is beyond the scope of this review, we may refer to two studies that assessed the performances of various non-coding variant prioritization tools in classifying sets of known deleterious HGMD variants. Each study compared a specific program developed by the authors to leading "state-of-the-art" algorithms. Fu et al. (Fu et al., 2014) showed that FunSeq2 has a better average prediction power compared to GWAVA and CAAD, while Shihab et al. (Shihab et al., 2015) showed that FATHMM-MKL outperformed GWAVA and CAAD in terms of accuracy. Due to the substantial number of recently developed methods, a full scale and independent comparative study would be valuable to provide consistent results and objectively identify the strengths and weakness of each tool.

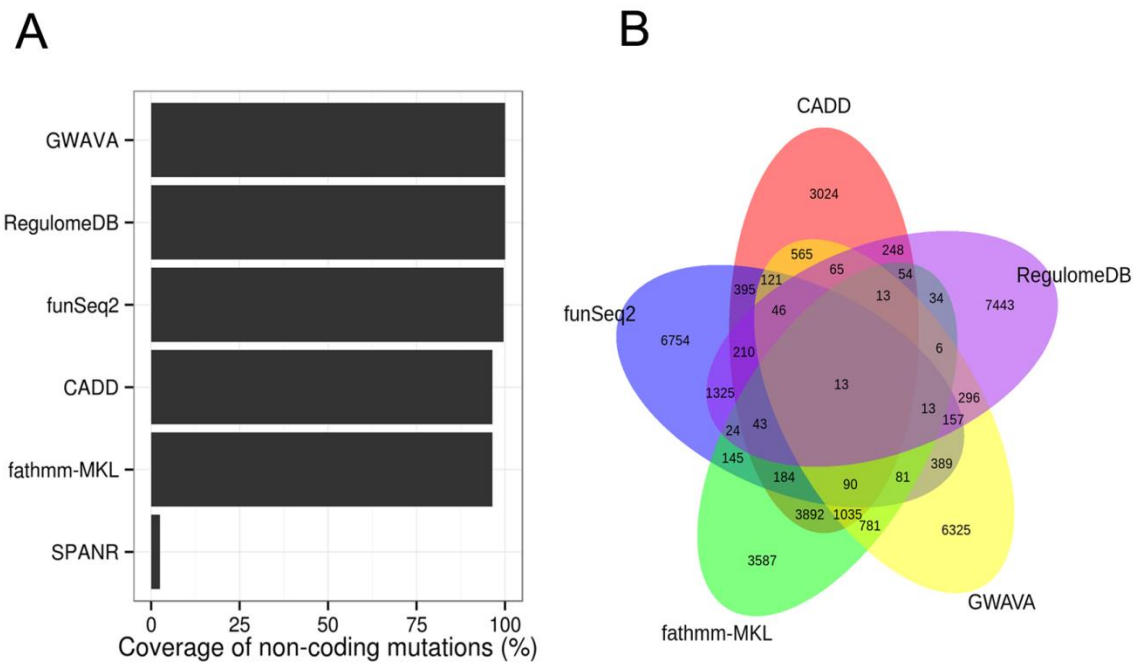

Figure 4. Comparison of six non-coding mutation scoring tools. A. Fraction of positions covered by each tool in a set of 874,325 non-coding variants. B. Overlap of the 10,000 top-scoring variants, using the 5 scoring tools with the larger prediction coverage (CADD, Fathmm-MKL, FunSeq2, GWAVA and RegulomeDB), from the 841,402 variants common to their prediction coverage. Five set of 10000 variants were randomly sampled from the 841,402 variants covered by five tools, the overlap was calculated accordingly. The sampling process was repeated 1000 times, the significance of overlap of 10000 high scoring variants was compared to the 1000 overlaps of random sampling variants.

Table 2. Summary of computational approaches for predicting the damaging effects of non-coding mutations

	Based on	**Machine learning**	**Cancer-specific**	**Web server, references**
Regulome DB	Overlap with functional elements	Empirical Scoring systems	No	http://www.regulomedb.org/(Boyle et al., 2012)
Funseq	Negative selection in general population recurrent cancer mutations	Empirical Scoring systems	Yes	http://funseq.gersteinlab.org/ (Khurana et al., 2013)
Funseq2	Negative selection in general population recurrence in cancer mutations	Empirical Scoring systems	Yes	http://funseq2.gersteinlab.org/ (Fu et al., 2014)
GWAVA	HGMD regulatory mutations, integrated genome annotation	Random Forest	No	https://www.sanger.ac.uk/sanger/StatGen_Gwava(Ritchie et al., 2014)
CADD	Deleteriousness,diverse genome annotation	support vector machine	No	http://cadd.gs.washington.edu/ (Kircher et al., 2014)
SPANR	RNA splicing model	Bayesian machine learning	No	http://tools.genes.toronto.edu/(Hs et al., 2015)
FATHMM-MKL	HGMD mutations, ten feature annotations (6 from ENCODE)	support vector machine	No	http://fathmm.biocompute.org.uk(Shihab et al., 2015)

1.5 Conclusion

The search for cancer drivers requires a reliable functional annotation of variants and adapted tools for analyzing the recurrence of deleterious variants across patients. The former requisite is particularly challenging in the non-coding genome. An active research community is developing tools for non-coding variant annotation and prioritization using a variety of methods ranging from empirical scoring scheme to machine-learning and elaborate hybrid frameworks. Due to the heterogeneity and complexities of these scoring tools, objective comparisons based on proper benchmarks using different sets of validated or probable

disease-causing variants are strongly required. Among multiple sources of possible improvement, the success of hybrid methods for scoring coding variants, and the widely divergent predictions by the non-coding tools suggest that combining outputs from different tools will significantly increase scoring accuracy for non-coding variants. A further challenge is to jointly consider this "functional" score and the heterogeneity of cancer specific mutation constraints in different genome areas. These potential enhancements suggest we can expect important reliability gains in non-coding variant prioritization in the near future.

As described above, there are a handful of computational tools used to evaluate the functional impact of non-coding mutations. However, certain limitations still exist for these prediction tools. For example, empirical scoring systems, such as RegulomeDB and funSeq2, cannot provide a precise measure of functional information for non-coding variants, while machine learning models, such as FATHMM-MKL and GWAVA, might be overfitted to a small set of HGMD disease mutations and show major ascertainment biases, and CADD doesn't take into account cancer mutation information in its scoring system. Moreover, although an increasing number of cancer-associated lncRNAs has been experimentally characterized, an efficient computational tool to prioritize cancer-driving lncRNAs is still missing, mainly owing to the sophisticated and diverse mechanisms by which lncRNAs act. Therefore, it becomes increasingly urgent and important to develop a scoring system that accurately measures the functional effect of non-coding cancer mutations and then injects this functional information into a computational program for the detection of non-coding drivers.

In the following studies, we hypothesized that purifying selection as measured by the fraction of rare SNPs in general population and mutation density (number of mutations /Mb) constraint are two important measures of functional impact of cancer mutations in the non-coding cancer genome. In order to functionally score non-coding mutations in cancer and eventually identify new cancer drivers, we took into account the dual selection forces acting on the tumor genome: (1) population and evolutionary constraints acting at germline level and (2) constraints resulting from the accelerated mutation background of the cancer tissue. To achieve this, we have developed two independent random forest models, referred to as SNP and SOM models. The SNP model predicts expected fraction of rare SNPs for any non-coding

region based on a combination of features, the SOM model computes the expected mutation density for each 1-Mb window with an array of feature types ranging from replication time, expression level, histone modifications to regulatory elements. The two models are capable of discriminating disease-associated variants from Clivariant and HGMD databases from a set of random control SNPs, strongly supporting our hypothesis. This study is the object of the following chapter.

Chapter 2 – Non-coding driver mutations

Results presented here are published in PLoS Computational Biology (Appendix 2)

A dual model for prioritizing cancer mutations in the non-coding genome based on germline and somatic events

.

LI J, Poursat MA, Drubay D**,** Motz A, Saci Z, Morillon A, Michiels S, Gautheret D**.** *PLoS Comput Biol***. 2015. 11(11):e1004583.**

Author contribution:

Jia LI was the main contributor to this study, he performed the whole experiment under the supervision of Professor Daniel Gautheret. Poursat Marie-Anne provided professional guidance as to the random forest model building and validation. Drubay Damien and Michiels Stephan gave statistical support to this work. Motz Arnaud was in charge of the preparation of figure 4 and supplementary figure 4. Professor Daniel Gautheret firstly wrote the manuscript, Jia LI, Saci Zohra, Morillon Antonin, Damien Drubay and Michiels Stephan gave their suggestion and comments. The paper was further revised by Daniel Gautheret and Jia LI together until the final acceptance.

2.1 Summary

Cancer cells undergo a mutation/selection process that resembles that of any living cell. Most mutations in cancer cell DNA occur in the so-called "non-coding" regions that represent 98.5% of the genome length. Pinning down which of these mutations contribute to the fitness of cancer cells would be important for identifying new "cancer drivers", which may in turn lead to future treatments. Unfortunately, predicting the impact of a non-coding DNA alteration remains extremely difficult. In this study, we analyze millions of non-coding cancer mutations and show cancer-specific mutational patterns can be used to predict non-coding regions that are preserved from mutations and may thus be important for cancer cell survival. Combining this information with population data, we propose a new scoring system that should help prioritize important non-coding mutations in future studies.

2.2 Introduction

Since the onset of cancer genomics, the search for cancer genes and cancer-causing mutations has largely focused on protein-coding genes and, more specifically, their coding exons, where the damaging effect of mutations is best understood. Among 572 human genes considered as cancer drivers (Futreal et al., 2004; D'Antonio and Ciccarelli, 2013), nearly all are protein-coding. However protein-coding regions only represent a tiny subset of the vast transcribed area composed of over 50,000 non-coding genes (Harrow et al., 2012; Iyer et al., 2015) and the introns and untranslated regions (UTRs) of mRNA genes. Even though a large part of the non-coding transcribed regions is probably non functional (Ulitsky and Bartel, 2013), analyses based on evolutionary conservation or allele frequencies in human populations (Ponting and Hardison, 2011; Ward and Kellis, 2012) estimate that 10 to 15% of the overall genome is under selection, that is 7-10 times larger than protein-coding regions.

Non-coding mutations may cause damaging effects in many distinct ways. They may alter RNA structure (Corley et al., 2015) or binding sites for proteins or other RNAs, such as splicing sites (Jolly et al., 1994) and microRNA target sites in 3' UTRs, or impact regulatory sequences in gene promoters and enhancers. A recent population genomics study estimates that there are in average 15 highly deleterious mutations in the non-coding DNA of any healthy individual (Khurana et al., 2013). This large source of potentially damaging mutation remains mostly untouched by cancer genomics. In-depth analysis of the mutational load in the non-coding fraction of the genome is needed for the comprehensive understanding of cancer progression, as well as for the identification of new cancer drivers and therapeutic targets.

Whole genome normal *vs*. tumor sequencing commonly reveals thousands to tens of thousands of somatic mutations (Alexandrov et al., 2013; Kandoth et al., 2013; Lawrence et al., 2013), scattered across all genomic areas. In coding regions the genetic code and aminoacid conservation rules provide a robust functional model for scoring mutational damage (Adzhubei et al., 2010; Ng and Henikoff, 2003). Similarly reliable tools are needed for non-coding regions in order to prioritize non-coding mutations and seek gene regions acquiring deleterious mutations at an unusual pace across a set of tumor samples. Several scoring systems for non-coding mutations already exist. The RegulomeDB system (Boyle et al., 2012) scores variants using an empirical metric based on their overlap with transcription

factor (TF) motifs, known TF binding site, chromatin marks or expression QTLs (eQTL) and thus is clearly centered on regulatory DNA variants. Other scoring models consider allele frequencies in human populations. Rare alleles are more often associated to reduced or lost gene activity than frequent alleles (Urban, 2005) and a high local ratio of rare to total SNP is indicative of purifying selection (Khurana et al., 2013; Chen and Rajewsky, 2006; Lomelin et al., 2010; Haerty and Ponting, 2013). Khurana et al. used SNP data from the 1000 Genome project (Clarke et al., 2012) to identify about 0,4% of the genome (12Mb) as sensitive to mutations and introduced an empirical scoring system (Funseq) to rate somatic mutations based on their presence in sensitive segments and overlap with known regulatory elements (Khurana et al., 2013; Fu et al., 2014). Likewise, the CADD system (Kircher et al., 2014) predicts the deleteriousness of non-coding mutations based on allele frequencies modeled using machine learning on a series of genome features. Recently, Ritchie et al. introduced a model for prioritizing non-coding variants based on databases of known disease-related mutations (Ritchie et al., 2014). The authors used machine learning to predict regions where disease-causing variants are most likely, using as explanatory variables functional features such as exon annotations, histone and other chromatin marks or transcription factor binding sites (TFBS). However useful, these models have limitations in that they are often directed towards the detection of regulatory elements (where 75% of disease variants have been located to date (Ritchie et al., 2014) and they only consider human mutations in the light of germline, evolutionary selection, meaning independently of a specific tissue or disease context. This latter point is especially important in cancer, where (1) most disease-inducing mutations occur somatically during the lifetime of an individual, and (2) these mutations may have different impacts when occurring in different tissues.

The availability of multiple whole genome sequence (WGS) data from tumors and matched normal tissue has revealed the extensiveness and singularity of cancer somatic mutations (Alexandrov et al., 2013; Kandoth et al., 2013; Lawrence et al., 2013). Cancer cells divide under their own set of selective constraints by which large regions of the genome can sustain high mutation rates while others seem relatively protected. This accelerated mutation rate is an important factor that may cause recurrent mutations in genome areas that are not necessarily related to cancer. Methods for scoring putative driver mutations now take such

effect into account (Lawrence et al., 2013).

Variation of the somatic mutation rates in different genome areas is by itself a rich source of functional information. Schuster-Böckler & Lehner (Schuster-Böckler and Lehner, 2012) related 45 functional features (mostly histone marks) to somatic mutation rates and observed that the major factor influencing mutation density was chromatin organization, marks of open chromatin being associated to a reduced SNV densities and marks of closed chromatin to higher densities. Cancer somatic mutations do not all cause cell death or tumor progression, but they may contribute to tumor heterogeneity which in turn facilitates the emergence of new clones capable of surviving micro-environmental changes and drug treatments (Podlaha et al., 2012). In this sense, the somatic mutation landscape can be considered as a model of accelerated evolution in which most mutations are neutral and a handful is under selection as beneficial to tumor progression.

A strong hypothesis guiding the present study is that, in order to prioritize non-coding mutations in cancer and eventually discover new cancer drivers, one should take into account these dual selection forces acting on the tumor genome: (1) population and evolutionary constraints acting at germline level and (2) constraints resulting from the accelerated mutation background of the cancer tissue. To this aim we developed two integrative models that use annotated genome features to predict germline or somatic mutation constraints at any genomic location. We compared the functional features that most influence each mutational regimen and analyzed the intersection of constrained regions predicted under each model. A new picture of the somatic mutational landscape emerges where regions under constraint in the germline may be subject to highly variable mutation rates in the tumor. We present evidence that low somatic mutation areas are functionally relevant and can be used as a powerful screen for prioritizing cancer-related non-coding mutations.

2.3 Results

We represent germline and somatic constraints acting on tumor genomes using two independent models, one for each mutational regimen, that we term the SNP model and the SOM model. For each model, we define a set of genome features, mainly from

UCSC/Ensembl genome annotation (Karolchik et al., 2014) and the ENCODE Project (Rosenbloom et al., 2013) and we use these features to predict the expected mutational constraint at any genome position. In the SNP model, the mutational constraint is expressed as a regional ratio of rare SNP, while in the SOM model it is expressed as a regional mutation density. We further describe each model below.

2.3.1 Scoring mutations with the germline (SNP) model

A high regional ratio of rare SNPs (*i.e.* SNPs with allele frequencies below 0.5 or 1%) is a hallmark of genome regions under negative / purifying selection (Chen and Rajewsky, 2006; Khurana et al., 2013; Haerty and Ponting, 2013). Figure 1A shows varying ratios of rare SNPs obtained from the 1000 Genome Project (Clarke et al., 2012) associated to known functional regions or "features"' (see Table S1 for each feature definition). Coding regions (CDS) clearly stand out as more constrained than non-coding regions in general. However, a number of non-coding elements also depart from the average genome signal, reflecting prior analysis of the 1000 Genome project data (Khurana et al., 2013). Regions under purifying selection (*ie.* with high rare SNP ratio) include evolutionary conserved regions, transcription factor binding sites, DNase I hypersensitive, early replicated and highly expressed regions. Inversely, we observed low ratios of rare SNPs in regions of strong GC-bias, high replication rate and evolutionary conserved RNA structures (ECS). Of note, this low ratio of rare SNP in ECS is in disagreement with the expected deleterious effect of mutations in functional RNA structures.

We developed a Random Forest (RF) model to predict purifying selection at any genome position based on the features present at this position. To this aim we associated every non-coding genome position to a vector of binary values describing the presence/absence of functional features at this location (see Table S1 and Methods). Following feature selection and cross-validation, we obtained a robust model associating any combination of 16 genomic variables to a predicted rare SNP ratio. A measure of importance of each feature's contribution to the RF model is shown in Fig.1B. Evolutionary conserved regions, promoters and conserved transcription factor binding sites are among the strongest contributors to rare SNP ratio, in line with previous studies (Clarke et al., 2012). Of note, the predictive value of a high recombination rate, which is associated to a low rare SNP ratio (Fig 5A), had not been reported before.

To evaluate how the SNP model alone can predict deleterious mutation in the non-coding genome, we compared the average scoring of one million random SNPs to that of non-coding variants from two distinct collections of disease-related mutations, the Clivariant (Landrum et al., 2014) and HGMD (Stenson et al., 2009) databases (Fig. 5C). Known clinical variants from either database have significantly higher scores by the SNP model than random variants (Wilcoxon P<2.2e-16 in both cases). Furthermore, scores in the SNP model are positively correlated to the density of disease-related SNPs (Fig 5D, r= 0.80 and 0.73, P=6.09e-08 and 3.15e-06 for Clivariant and HGMD, respectively), which confirms the capacity of the SNP model to identify non-coding regions where mutations are more likely to be disease-related.

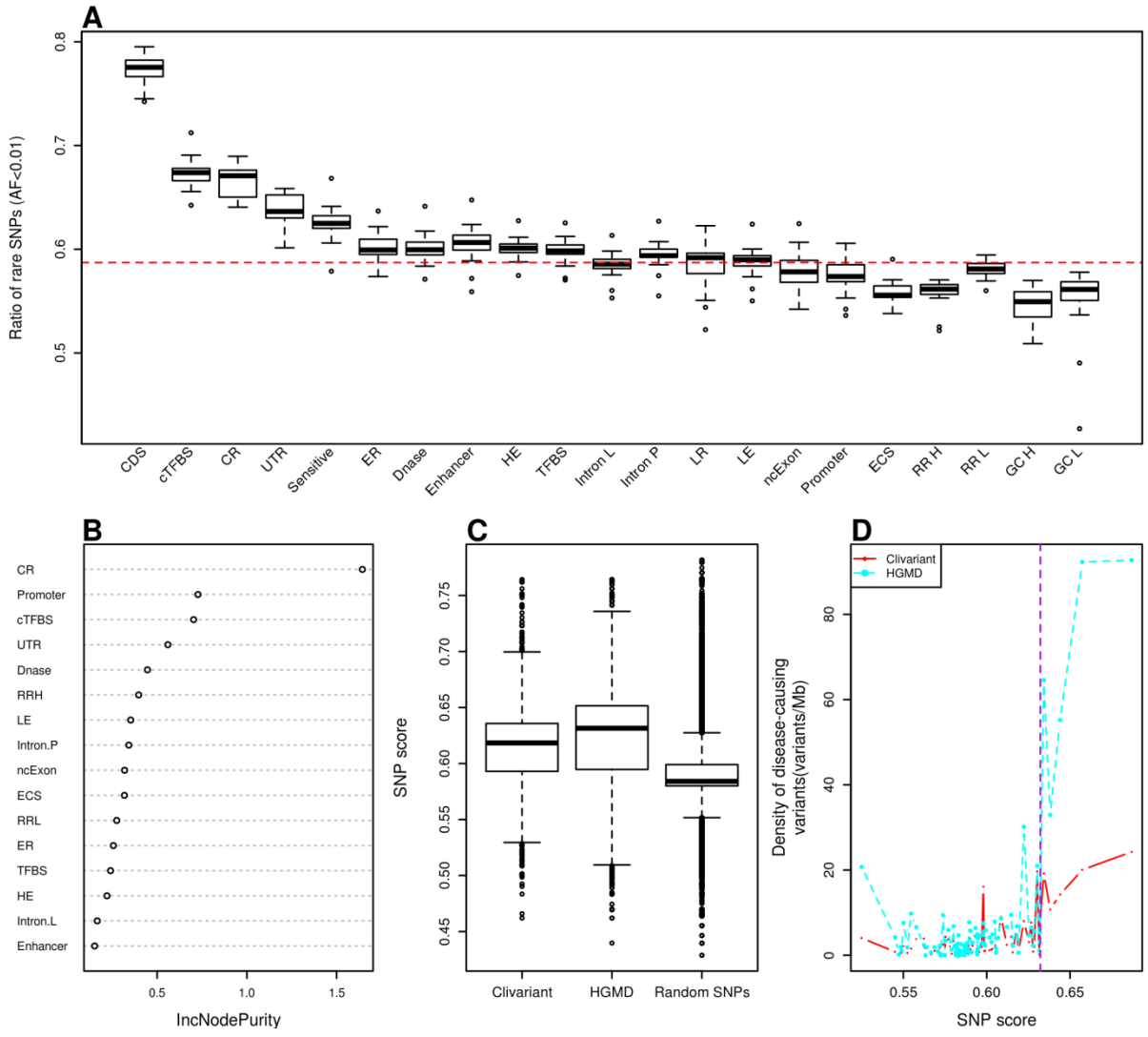

Figure 5. Construction of the rare SNP model. **A.** Fraction of rare SNPs (allele frequency <0.01)

according to different genome features (see Table S1 and Methods for feature details). Each box shows rare SNP fraction across all human chromosomes, except chr. Y. CDS: coding sequence; cTFBS: conserved transcription factor binding site; CR: evolutionary conserved region; UTR: untranslated region; Sensitive: region with high rate of rare SNP defined in (Khurana et al., 2013), ER/LR: early and late replicated region; DNase: DNase I hypersensitive site; HE/LE: high and low expressed region; Intron L/Intron P: intron of lncRNA/of protein coding gene; ncExon: non coding exon; ECS: evolutionarily conserved structure; RR H/RR L/GC H/GC L: high recombination rate, low recombination rate, high GC content and low GC content regions. The red dotted line represents the average fraction of rare SNPs across the genome. **B.** Feature importance as measured by IncNodePurity. We only show here features that passed feature selection. **C**. Distribution of SNP scores for random SNPs and for clinical variants from the Clivariants and HGMD databases. Random SNPs here are a set of 1M random intergenic SNPs from the 1000 Genome project. **D**. Correlation of SNP scores with densities of disease-causing variants. Genome positions were sorted by SNP score and split into 20 Mb intervals. The plots show the average SNP score and density of disease-causing variants for each interval. The purple dotted line shows cutoff used for defining high SNP score thereafter.

2.3.2 Scoring mutations with the somatic (SOM) model

The tumor mutational landscape results from the combined action of multiple factors including mutagenic agents, accelerated cell division, impairment of DNA replication/repair pathways and resistance to treatment (Lawrence et al., 2013). The tumor genome is thus subject to a set of constraints that are quite distinct from those acting in the germline. To analyze these constraints, we collected somatic mutation data from whole genome sequencing of liver cancer (N=88 patients), chronic lymphocytic leukemia (CLL) (N=28), lung adenocarcinoma (N=24) (Alexandrov et al., 2013) and melanoma (N=25) (Berger et al., 2012). We analyzed mutation densities for the above genomic features and for tissue-specific features such as histone marks, early/late replicated regions and transcript abundance obtained from tissue-matched Encode cell lines (Rosenbloom et al., 2013) (Table S2). Results are shown in Figure 6A, S1A, S2A, S3A. Protein-coding sequences (CDS) harbor relatively low somatic mutation densities compared to introns (intron.P) and intergenic regions in all four cancer types, consistent with higher functional constraints in CDS, as observed in the SNP

model. However, other features reveal a quite different pattern. Evolutionary conserved regions, cTFBS and UTRs that were all under strong selective constraints in the germline model present highly variable mutation densities in tumors, with densities ranging from low (CDS level) to high (intergenic level), and no consistent pattern from tumor to tumor (Fig 6A, S1A, S2A, S3A). Certain features, however, present marked and consistent mutational patterns across all four tumors. For instance, we observed an obvious trend for accelerated mutation rates (higher density) in regions of repressed chromatin marks (H3K9me3), late replication (PCgene.late, lncRNA.late), low transcript expression (PCgene.LE, lncRNA.LE) and low GC (GC L). Conversely, we observed consistently reduced mutation rates in regions of active chromatin marks (H3K4me1-2-3, H3K79me2, H4K20me1), early replication (PCgene.early, lncRNA.early), high transcript expression (PCgene.HE, lncRNA.HE) and high GC (GC H). The general trends in feature-wise mutation densities largely reflect prior findings based on smaller datasets. Schuster-Bockler and Lehner (Schuster-Böckler and Lehner, 2012) observed strong correlations between chromatin states and mutation densities in tumors, with repressive marks linked to higher mutation rates, possibly due to deficient DNA repair in these regions. Mutation density is also known to correlate positively with late replication (Hodgkinson et al., 2012; Lawrence et al., 2013; Woo and Li, 2012) and negatively with recombination rate (Schuster-Böckler and Lehner, 2012) and RNA expression level (Lawrence et al., 2013; Pleasance et al., 2010).

To model the mutational constraints acting on the tumor genome, we developed a second RF model, referred to as the SOM model, which predicts somatic mutation densities (the response variable) at any genome position based on the presence of cell-specific and generic genome features. We built one SOM model for each of the four above cancer types. Due to the large number of features in the SOM model and limited number of somatic mutations in the training sets, we computed feature coverage or average values (see methods) on successive 1Mb regions and trained the RF model based on the resulting vectors. After feature selection and robustness testing by cross-validation, the SOM model enabled reliable prediction of somatic mutation density at any genome location for each cancer type (see Methods). Fig 6B, S1B, S2B, S3B show the importance of features in the SOM models.

RNA expression levels turned out to be relatively weak predictors of mutation density,

whereas replication time and histone marks in general are the predominant features determining somatic mutation density in all cancer types. However we observe significant differences between cancers. For instance the H3K36Me3 mark is an important predictor of low mutation density in melanoma and lung cancer, not in CLL or liver cancer. Also, CTCF binding sites are strong predictors of low mutation density in CLL and not in other cancer. Altogether this indicates that each somatic model predicts a cancer-specific mutation profile with distinct regions of high and low mutation densities.

Under a neutral evolutionary model, somatic mutations should freely accumulate in regions that do not impact tumor fitness, thus regions of elevated tumor densities (high SOM score) should be considered as generally irrelevant to fitness, while regions that are relatively preserved from somatic mutations (low SOM score) are potentially the most interesting as they could reveal purifying selection occurring at the tumor level. One way to test this hypothesis is to relate low mutation regions and the occurrence of known disease mutations. Fig 6C, S1C, S2C, S3C show that non-coding disease mutations from the Clivariant and HGMD databases have significantly lower SOM scores than evolutionarily neutral SNPs (Wilcoxon $P<2.2e-16$ in all cases). Furthermore, the SOM score of different genome regions is inversely correlated to the density of disease causing variants in these regions (Fig 6D, S1D, S2D, S3D) ($r = -0.47$ to -0.94, $P = 0.01$ to $8.61e-14$) suggesting that genome regions spared from somatic mutations are functionally relevant to disease progression.

To further assess the value of SOM score as an indicator of selection, we mapped the genome positions with lowest SOM scores onto the different genome features and measured the relative enrichment for low SOM score positions within each feature (Fig. S4A). Expectedly, features that were part of the SOM model are significantly enriched or depleted in low SOM scores. However, 5' and 3' splice sites, two features that were not part of the model, show a much higher coverage by low SOM score regions than intronic regions, which indicates functional non-coding elements tend to attract fewer somatic mutations, as expected under a negative selection model. This effect is also observed in lncRNA, consistent with the higher conservation of splice junctions in this class of genes (Nitsche A, Rose D, Fasold M, Reiche K, 2015). Conversely, features enriched in high SOM scores (Fig. S4B) predominantly correspond to silent regions (intergenic, centromeres and telomeres). In summary low SOM

score positions tend to colocalize with functional elements and correlate with disease-causing mutations, suggesting the SOM model could be a significant, independent source of functional information on non-coding regions.

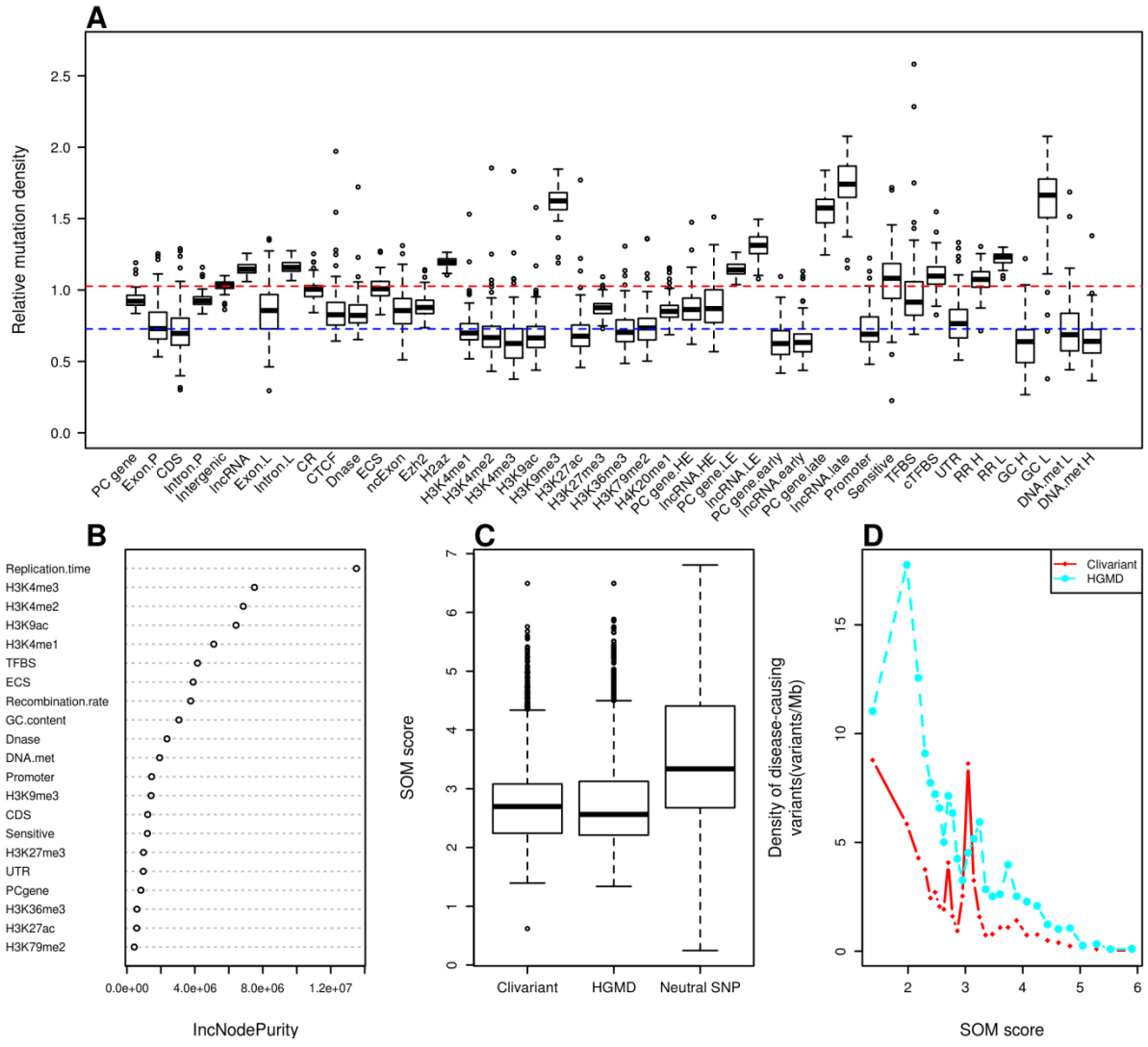

Figure 6. Construction of the Somatic Mutation (SOM) model for liver cancer. **A**. Relative density of somatic mutations from whole genome sequences of 88 liver tumors (Alexandrov et al., 2013), associated to different genome features (see Methods for feature details). Mutation density is normalized so that the whole genome average has a mutation density of 1. PC gene: protein coding gene; CDS: coding sequence; Exon.P, Intron.P, Exon.L,Intron.L are exon and intron of protein coding gene and lncRNA respectively; CR: conserved region; DNase: DNase I hypersensitive site; ECS: evolutionarily conserved structure; ncExon: non-coding exon; PC gene.HE, LncRNA.HE, PC

gene.LE and LncRNA.LE are high expressed and low expressed protein coding gene and lncRNA; PC gene.early, LncRNA.early, PC gene.late and LncRNA.late are early and late replicated protein coding gene and lncRNA; cTFBS: conserved transcription factor binding site;RR H,RR L,GC H,GC L,DNA.met H and DNA.met L are 1-Kb windows with high recombination rate (> 4.0), low recombination rate (< 0.5), high GC content (GC % > 50%), low GC content (GC%<30%), high DNA methylation (average value > 0.7245) and low DNA methylation (average value < 0.4062) respectively; Blue and red dotted lines: base lines showing average values for CDS and intergenic regions, respectively; **B:** Feature importance as measured by IncNodePurity. We only show here features that passed feature selection. **C**. Distribution of SOM scores for neutral SNPs and for clinical variants from two disease-causing variants databases Clivariant and HGMD. Neutral SNPs here are SNPs from the 1000 Genome project with allele frequency higher than 0.01, SOM scores predicted by the random forest model were divided by the number of patients. **D**. Correlation of SOM score with densities of disease-causing variants. Genome positions were sorted by SOM score and split into 100Mb intervals. The plots show the average SOM score and density of disease-causing variants for each interval. The purple dotted line shows cutoff used for defining low SOM score thereafter.

2.3.3 Towards an integrated model for germline and somatic mutations

Analysis of germline and somatic mutations suggests that each mutational regime carries valuable independent information about selective forces acting in a tumor. We thus questioned whether combining SNP and SOM information at each genome position may lead to improved mutation prioritization in cancer.

To assess the benefits of the joint model for scoring disease mutations, we measured disease variant densities in different areas of each tumor spectrum using the above cutoffs (Table S3, Fig S6). If we intersect high-SNP and low-SOM regions, the resulting genome area shows a greater enrichment in disease variants than either region taken independently ($P<2.2e-16$ for all four cancers). Therefore we argue that integrating germline and somatic mutational models provide a better system for prioritizing damaging mutation than any model used independently.

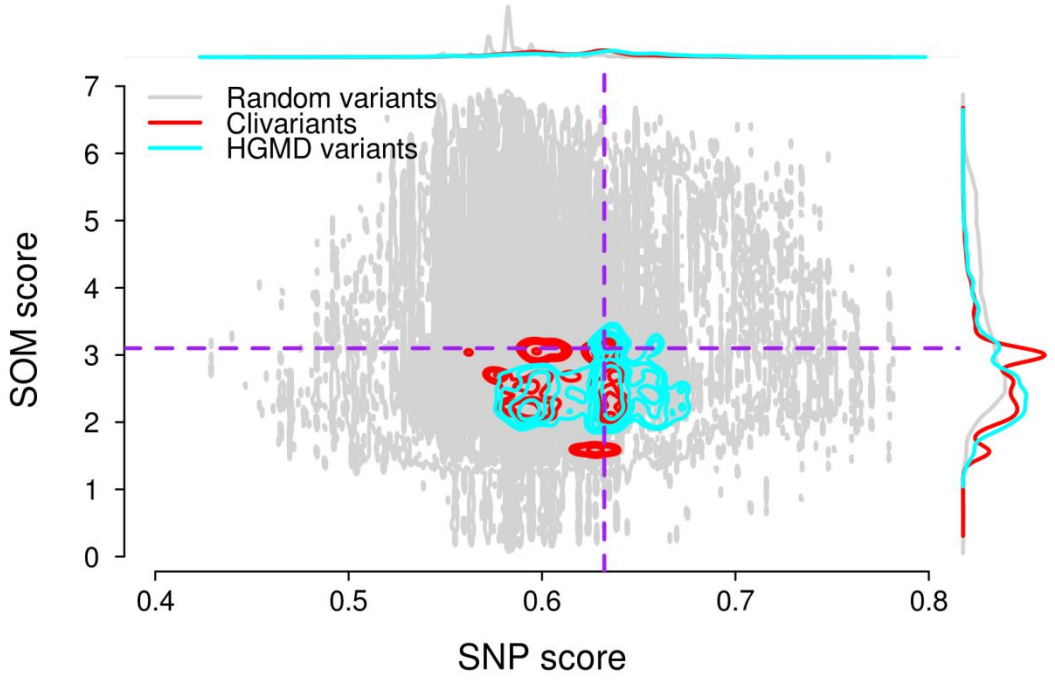

Figure 7. Relationship between SNP and SOM scores in liver cancer. Contours show densities of positions with the corresponding SNP and SOM scores. Grey dots: 1 million random genome positions; cyan contour: HGMD disease-causing variant positions; red contour: Clivariant positions. The top and right curves show marginal distributions of SNP scores (top) and SOM scores (right) for random genome positions, HGMD and Clivariant disease-causing variant positions. Dotted lines define cutoff values for hypomutated/hypermutated regions. SNP score cutoff=0.63 (98.16Mb above cutoff), SOM score cutoffs = 3.10 variants/Mb, defining areas below cutoff of 55.67 Mb, in liver cancer. Hypomutated regions defined by both cutoff correspond to ~56Mb in liver cancer type.

Hypomutated positions are significantly over-represented within splice junctions, UTRs and different classes of cancer genes. We mapped predicted hypomutated positions on different genome features and gene types (Fig 8). As expected, functional features of protein-coding genes such as intron junctions and UTRs are strongly enriched for hypomutated positions (Fig 8A). Similar trends are observed in lncRNA genes. Both lncRNA introns and exons are generally depleted for hypomutated regions (Fig 8), in line with poor selective pressure in lncRNA overall. However, lncRNA splice sites are slightly, albeit significantly, enriched in hypomutated regions, consistent with previous studies showing increased purifying selection

at lncRNA splice sites (Nitsche A, Rose D, Fasold M, Reiche K, 2015).

We then compared hypomutated position enrichment in cancer vs. non-cancer genes. Cancer-protein-coding genes and cancer-related miRNAs are enriched for hypomutated regions compared to their non-cancer counterparts (Fig 8B, Table S4). This result suggests an elevated protection from somatic and germline mutations in cancer miRNAs and in the introns and UTRs of known cancer genes (we remind our analysis only considers the non-coding part of genes). However, we did not observe a significant enrichment for hypomutated regions in our short list of cancer-related lncRNAs (N=25). Complete lists of protein-coding, lncRNA and miRNA genes with their fraction of hypomutated positions are provided as suppl. files. Notable cancer genes with high fractions of hypomutated positions include PIM1 and MED12, with respectively 34% and 32% of their non-coding length that is hypomutated. Among cancer miRNAs, miR-1 and miR-574 are both covered almost completely by hypomutated positions.

Interestingly, genes with high fractions of hypermutated positions are more divergent between cancer types than genes with high fractions of hypomutated positions (Fig S7), suggesting areas of high mutation density are largely cancer-specific, while areas of low mutation density tend to locate in the same functional regions of the genome. GO-term biases in these gene sets are significant only for genes enriched for hypermutated positions in liver cancer and CLL, and involve transcription regulation functions (Table S5).

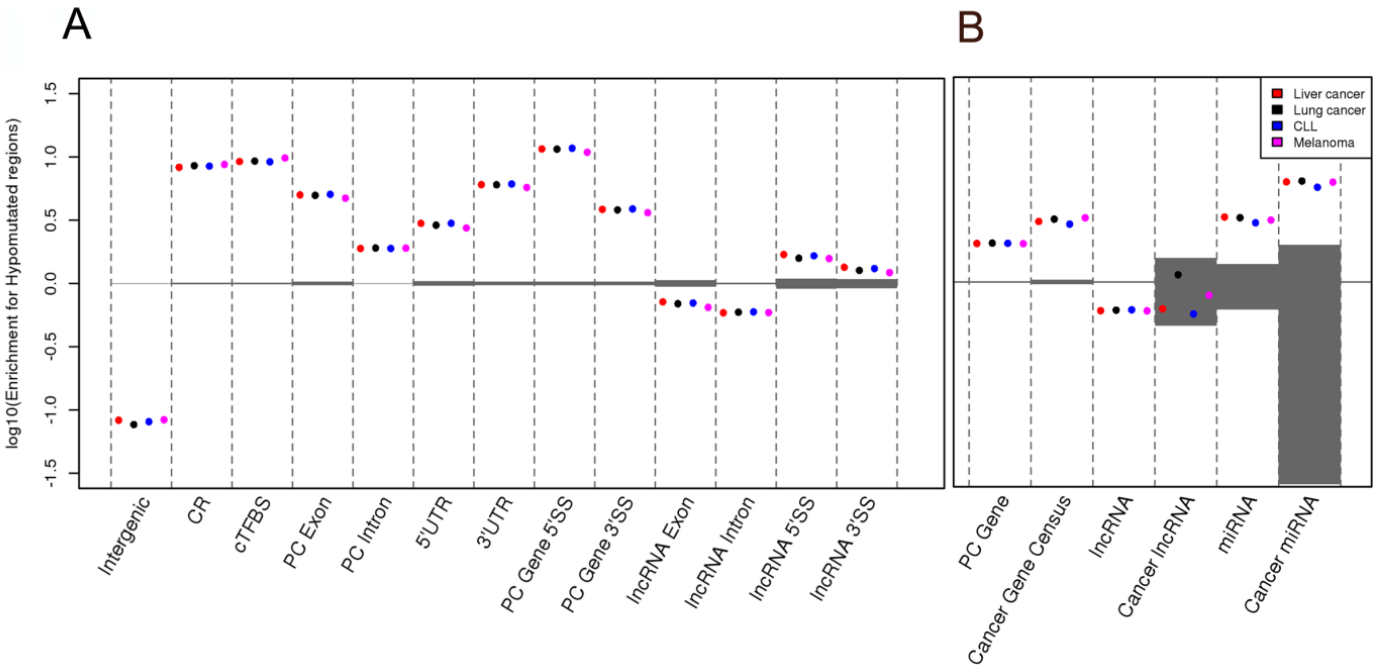

Figure 8. Enrichment for hypomutated positions within different genome features (A) and gene classes (B). Positive values indicate enrichment, negative values indicate depletion. Hypomutated (high SNP, low SOM) positions were mapped onto genome features (A) or genes from three different classes (Protein-coding, lncRNA, miRNA) (B). For each feature or gene class, enrichment for hypomutated positions was computed as explained in Methods. As hypomutated positions are cancer-specific, different results are obtained for each cancer class (colored dots). Shaded grey areas show enrichment ranges obtained from 1000 random permutations (see Methods).

2.4 Discussion

We introduced novel computational models to assess mutational constraints in the non-coding genome based on the presence of functional features. We trained a model on germline SNP data to predict rare SNP ratio at any genome site, and we trained four cancer-specific models on tumor data to predict somatic mutation densities. These models thus provide two independent measures of mutational constraints that are both relevant to the analysis of non-

coding regions in the cancer context. Furthermore, the feature-based model construction enabled us to analyze the contribution of each feature to the germline and tumor mutation landscape and to characterize the main differences between the two mutational regimens.

A major point we want to highlight in this study is that combining germline and somatic data provide an improved definition of non-coding regions that are sensitive to mutation in cancer cells. To illustrate this point, we extracted genome areas combining a high rare SNP ratio and a low somatic mutation density and showed these combined criteria are a better predictor of disease causing mutation than rare SNP ratio or somatic mutation density considered independently.

Distinctly from current models that consider somatic mutation only as a corrective mean to avoid overpredicting deleterious mutations in highly mutated regions (Khurana et al., 2013; Lawrence et al., 2013; Fu et al., 2014), our approach thus considers somatic mutations on a par with evolutionary mutations, that is as a criterion to tell apart genome positions that are neutral (highly mutated) or under purifying selection (lowly mutated) in the tumor genome. We remind that prevalent forces shaping the tumor mutation landscape are the combined actions of mutagens and the DNA repair machinery on differentially accessible genome regions (Guttman et al., 2011; Schuster-Böckler and Lehner, 2012; Watson et al., 2013). Therefore, if functional areas are relatively spared from mutation, this is mostly not as a result of purifying selection, but because they are under the closer watch of DNA repair systems. Hence the somatic model can be viewed primarily as a way to discard regions sustaining accelerated mutations. However, we showed that hypomutated regions were enriched in functional elements such as splice junctions, which suggests purifying selection may occur as well.

We are aware of the limited accuracy of somatic models when these are trained over tumors with low mutation rates and/or few available whole genome datasets. Currently, there are far fewer mutations to learn from in the tumor dataset than in the human polymorphism dataset (aggregate mutation densities in the present cancer datasets ranged from 20 to 600 mutations per Mb, vs. >12,000 SNP per Mb in the 1000 Genome data). This limits our ability to observe small-scale variations in mutation density. We expect that the fast accumulation of whole

tumor sequences will improve model accuracy within each cancer type and provide independent validation of our approach on other tumor classes. Another potential limitation in SOM models is the use of expression and epigenetic features from cell lines as a proxy for cancer tissues. This should also improve in the future as such information is acquired from primary tumor tissues.

A key outcome of our study is a new approach to prioritize non-coding variations for cancer driver search. Our models predict mutational constraints at a genome position based on generic features, that is, largely independently of the actual mutations observed at this specific location. Therefore, a locus may be predicted as hypomutated by the model and yet turn out to sustain recurrent mutations across patients. Such a locus should then be prioritized as a candidate driver. Such analyses will be natural extensions of the present study.

Although cancer research now acknowledges the importance of non-coding drivers, the search for cancer-related mutations has focused on regulatory elements such as promoters and enhancers as the key non-coding elements (Khurana et al., 2013; Ritchie et al., 2014). The realization that nearly 60,000 lncRNAs are expressed, often specifically, in tumoral genomes, many of them harboring potential disease causing mutations (Iyer et al., 2015), combined to the regulatory roles played by many lncRNAs (Forbes et al., 2011a) indicate that cancer driver search should also encompass those larger transcribed regions. Even if only 10% of lncRNAs are functional by conservative estimates (Ulitsky and Bartel, 2013), this corresponds to a much larger genome area than known regulatory elements. Currently, the search for cancer genes in these non-coding RNAs is driven by expression signature analysis. We show here that the analysis of germline and somatic mutational regimen is an important alternative that may lead to the identification of cancer-driving elements in ncRNA genes, as well as in the non-coding fraction of mRNA genes.

2.5 Materials and Methods

2.5.1 Human polymorphism, mutation and disease data

Human polymorphism data comprising 38,248,779 SNPs were downloaded from the 1000 Genome project pilot 1 (Clarke et al., 2012) (http://www.1000genomes.org). The data set

contains SNP data from 2500 individuals from about 25 world populations. SNPs with allele frequency lower than 0.01 were defined as rare, other SNPs were considered neutral.

Somatic variants were collected from whole genome sequencing of paired cancer and normal tissues, obtained from two studies: 2,011,261 variants from 25 melanoma patients (Berger et al., 2012), 1,845,976 from 24 lung adenocarcinoma patients, 881,136 from 88 liver cancer patients and 59,993 from 28 chronic lymphocytic leukemia (CLL) patients (Lawrence et al., 2013). Variants described as "substitution" or "indel" were both collected and are referred to collectively as mutations in the text.

Curated disease-related variants were obtained from the Clivariant (Version 2014/03/03, 55,689 variants) (Landrum et al., 2014) and HGMD (Version 2014/04/14, 166,768 variants) databases (Stenson et al., 2009). After exclusion of coding positions we used 13,108 HGMD and 6045 Clivariant mutations.

Lists of cancer genes for Fig. 8 were obtained as follows: protein-coding cancer genes are from the Cancer Gene census, available from COSMIC release V71 (http://cancer.sanger.ac.uk/cancergenome/projects/census/) (Forbes et al., 2011a); cancer-related lncRNAs are 27 mammalian long non-coding transcripts identified from our literature search as experimentally associated with different cancer types (Table S6); cancer miRNAs are from the miRCancer database (Andersson et al., 2014).

2.5.2 Uniform genome-wide features

Uniform features used in all figures and models are summarized in Table S1. Human genome annotation (protein-coding and lncRNA genes, exons, introns, CDS, UTRs, non-coding Exons (ncExon) was obtained from Gencode V7 (Harrow et al., 2012). We defined as intergenic those regions covered by neither a protein-coding gene (including introns) nor an lncRNA. We defined as 5’ and 3' splice sites intron regions spanning the first 10 nt on the 5' side and the last 50 nt on the 3' side. GC contents were computed directly from the HG19 human genome assembly. We defined 1kb regions with > 50% GC as high GC and 1kb regions with < 30% GC as low GC. For the SOM model, GC contents were computed over 1Mb windows.

Promoters, defined as regions of 2.5kb from transcription start site (TSS), are from the

Gerstein lab (http://funseq.gersteinlab.org/data) (Khurana et al., 2013). Enhancers are from the Atlas of active *in vivo*-transcribed enhancers, collected based on FANTOM5 CAGE data from multiple tissues and cell lines (Karolchik et al., 2014). TFBSs combine all transcription factor binding sites from more than 30 Encode cell lines (Rosenbloom et al., 2013). Conserved TFBS (cTFBS) are from the UCSC tfbsConsSite track established from human/mouse/rat alignment (Smith et al., 2013).

"Sensitive regions" are defined in the Khurana et al. study of genome regions under purifying selection as the 0.4% genome fraction with highest enrichment in rare SNPs (Khurana et al., 2013). Evolutionarily conserved regions (CR) are from the UCSC 46 mammalian genome alignment (Phastcons score >177) (Smith et al., 2013). Evolutionarily conserved structures (ECS) are RNA secondary structures predicted using comparative structure prediction algorithms based on multiple genome alignments (Altshuler et al., 2010). DNase I hypersensitive sites (DNase I) from 125 combined ENCODE cell lines were obtained directly from the UCSC web site (Rosenbloom et al., 2013).

We defined early and late replication regions using the ENCODE ‘Repli-seq' track (http://genome.ucsc.edu/ENCODE) that provides signals for cell cycle fractions G1b, S1, S4, G2 in different cell types (Rosenbloom et al., 2013). For each protein-coding or lncRNA gene, we computed the early-to-late (E/L) ratio as (G1b+S1)/(S4+G2) averaged over the gene length. Early and late replicated genes denote genes or lncRNAs with an E/L ratio > 1 or < 1 for all 10 cell lines respectively: Gm12878, Hela3, Hepg2, Mcf7, Imr90, K562, Bg02es, Huvec, Bj and SK-N-SH.

Expression levels were calculated using number of reads per kilobase per million reads (RPKM). We defined as High Expression (HE) genes those with RPKM > 20 in any of the 27 Encode cell lines (Rosenbloom et al., 2013), corresponding to the top 6% of protein coding genes for a single Encode cell line.

Recombination rates (RR) are from the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) (Breiman, 2001). As every genome position did not have an associated RR, we averaged HapMap RR values over 1kb windows. High replication rate (RRH) and low replication rate (RRL) regions were defined by an average replication rate above 4.0 or below

0.5, respectively.

2.5.3 Tissue-specific features

RNA expression levels, transcription factor binding sites (TFBS) and maps of histone modification marks were acquired from UCSC ENCODE tracks (Rosenbloom et al., 2013) for each cell type: Hepg2, A549, K562, Nhdfad (Table S2). Replication timings were acquired from UCSC ENCODE tracks for cell lines Hepg2, lmr90, K562, Bg02 (Table S2).

To define high expression and low expression genes, expression levels were measured for a single randomly selected cell line from the same tissue for each independent protein coding gene and lncRNA. RPKM values above 20 and below 0.25 defined high (PCgene.HE, lncRNA.HE) and low expression genes (PCgene.LE, lncRNA.LE), respectively.

Replication timings were defined for each protein-coding gene and lncRNA using the same E/L calculation as above. Genes with an E/L ratio > 1 were considered early replicated (lncRNA.early, PCgene.early), genes with an E/L ratio < 1 were considered late replicated (lncRNA.late, PCgene.late).

DNA methylation data were obtained from TCGA database (http://cancergenome.nih.gov/) for cancer types liver hepatocelluar carcinoma, lung adenocarcinoma, acute myeloid leukemia and skin cutaneous melanoma. Average DNA methylation value was computed for each methylation site across multiple patients, undefined values were replaced with mean and then we averaged DNA methylation over non-overlapping 1Kb and 1Mb windows, 1Kb windows which have mean DNA methylation values greater than 0.7245 and less than 0.4062 were defined as high (DNA.met H) and low (DNA.met L) DNA methylation windows respectively.

2.5.4 Rare SNP model

A random forest (RF) is an ensemble of multiple decision trees computed from separate bootstrap samples of the training data and feature set (Breiman, 2001). We developed the germline RF model (SNP model) to predict the density of rare SNP at any genome location based on 14 distinct features (Table S1). The response variable was the local ratio of rare SNP (number of rare SNPs /total number of SNPs) obtained from the 1000 Genome Project.

A matrix of 44130 rows was formed after removal of those combinations in coding regions,

each row representing one type of combination of features that can be observed throughout the non-coding genome. Feature selection was performed with the R VSURF package (Genuer et al., 2012), resulting in elimination of GC which is G or C base for each nucleotide and late replicated regions, 18656 combinations of the remaining 16 features. 2502 combinations of 16 features containing 99.49% of SNPs and 99.50% of human genome were used to train the model after removal of the combinations of size smaller than 10Kb. The RF model was produced using the R randomForest package. The SNP score was predicted with the 16 selected features for each combination of feature in the non-coding genome. Model calibration and cross validation are presented in Supplementary methods. Variable importance was estimated using node purity, which measures the decrease in tree node purity that results from splits of a given variable.

2.5.5 Somatic mutation model

The somatic (SOM) RF model was built using as predictors the 16 uniform and 17 tissue-specific features described in Table S1 and S2, and as response variable the local density of somatic mutation across all tumors in the cancer type under study. Due to the relatively sparse somatic mutation data, model fitting was performed using continuous variables measured for genome windows as explained below.

Features ncExon, introns of lncRNAs and PC genes, CR, cTFBS, UTR, Promoter, GC contents and the various histone marks were expressed as the number of nucleotides covered by the feature within each successive 1Mb window. Features recombination rate, DNA methylation, replication time and expression level were computed for each successive 1Mb window as follows. To obtain expression levels for 1Mb windows, RNA-seq reads from each cell lines (3 samples/cell line) were counted, and the length of exons from Gencode annotation was calculated, then, average expression level was calculated as RPKM. Replication time in the SOM model was the average E/L ratio computed as above for each 1Mb window. Recombination rate and DNA methylation were averaged over non-overlapping 1-Mb windows across the genome.

The SOM model used cancer mutation density as the response variable and the 33 genomic features (32 for lung cancer) as predictor variables. A matrix of 2846 rows was formed, of

which each row represents a 1-Mb window and columns contain values of genomic features and response variable. For model fitting, we discarded genome regions with poor annotation or biased mutation information. This included any 1Mb window overlapping a telomere, centromere, stalk, pericentromere, or with 100% undefined bases, and the entire Y chromosome due to ploidy bias (total: 224.3Mb). All predictor values were plus one and log scaled.

The RF regression model was constructed with the R randomForest package as above. Feature selection was performed with the R VSURF package (Genuer et al., 2012). Model calibration, robustness testing/cross validation of the SOM models are presented in supplementary methods. For SOM score prediction, we used the same 1-Mb window strategy as in model building, however, the 1Mb-windows were slided across the human genome with a step size of 1Kb, in order to extrapolate to regions not used in model building. 1Mb windows with annotation or mutational biases were excluded as in model training, resulting in 2,832,687 overlapping 1Mb window annotations. The SOM score was predicted using selected features for each 1Mb window and averaged on a 1 Kb window scale.

2.5.6 Enrichment analysis

Enrichment for hypomutated positions within different feature classes (Fig 8) was measured as the odds ratio:

$$enrichment = \frac{\left(\frac{Hf}{Sf}\right)}{\left(\frac{Hg}{Sg}\right)}$$

Where Hf = #hypomutated positions within feature, Sf = total size of feature, Hg = #hypomutated positions in whole genome, Sg = total size of genome. The significance of enrichment or depletion was evaluated using a permutation test as follows: a set of positions of same size as the hypomutated region (ie. 56Mb) was randomly sampled from the whole genome 1000 times, and in each random sample, enrichments were calculated for each feature class. The distribution of enrichment values from the 1000 random samples is shown as shaded areas in Figures. Only observed enrichments outside these areas are considered significant. Enrichment for other types of positions (hypermutated, low SOM score etc.) was evaluated similarly.

Chapter 3 –LncRNAs and cancer

Author contribution:

Jia LI firstly wrote the Chapter 3, Daniel Gautheret gave his suggestion and comments and further revised this section.

3.1 Introduction

Cancer is the second leading cause of deaths in USA, about 1,658,370 new cancer incidences and 589,430 mortalities are estimated to occur in USA in 2015(Facts, 2015). Cancer is characterized by uncontrolled growth of malignant cells. Causes of cancer are complex and diverse, ranging from external factors such as mutagenic agents and infectious organisms to internal factors such as inherited mutations and immune deregulation (Gutschner and Diederichs, 2012). In 2000, Hanahan and Weinberg proposed 6 critical capabilities that cancer cells possess to enable the malignant transformation, including sustaining proliferative signaling, evading growth suppressors, enabling replicative immortality, activating invasion and metastasis, inducing of angiogenesis and resisting cell death (Hanahan and Weinberg, 2011). Detection of driver genes critical to these events is a consistent goal in cancer genomics. Multiple bioinformatic tools have been developed to discriminate cancer-driving genes from background genes, such as MutSigCV (Lawrence et al., 2013) and MuSiC (Dees et al., 2012) which search for recurrently mutated genes across a cohort of cancer samples and Oncodrive-fm (Gonzalez-Perez and Lopez-Bigas, 2012) which determines driver genes accumulating mutations with high function effect. Up to now, 547 driver genes have been identified and annotated in COSMIC database (Forbes et al., 2011b).

LncRNAs are a class of mRNA-like transcripts ranging from 200 bp to 100 kb, which lack significant open reading frames and are not translated into proteins. A recent compendium found 58648 lncRNAs in the human transcriptome (Iyer et al., 2015). LncRNAs are mostly two-exon transcripts and preferentially localized in chromatin and nucleus. They show lower expression and higher tissue specificity as compared to protein coding genes (Derrien et al., 2012). According to their genetic relation with protein coding genes, lncRNAs can be classified into five main categories: sense and antisense lncRNAs which are located in a transcript on the same or opposite strand, respectively, bidirectional lncRNAs whose expression and neighboring transcripts on the opposite strand are transcribed in close genomic proximity, intronic and intergenic lncRNAs which are derived from intronic and intergenic regions of transcripts respectively (Ponting et al., 2009).

LncRNAs were initially thought to be spurious transcriptional noise due to low RNA polymerase fidelity. In recent years, accumulating evidences have shown that lncRNAs are pervasively transcribed throughout eukaryotic genomes and involved in a wide range of physiological processes, such as imprinting (Jeon et al., 2012), epigenetic regulation (Mattick et al., 2009), apoptosis and cell cycle control (Wapinski and Chang, 2011), transcriptional (Orom et al., 2010) and translational regulation, splicing, cell development and differentiation (Clark and Mattick, 2011) and aging (Rando and Chang, 2012).

Despite their lack of protein-coding capability, many lncRNAs are suspected to harbor biological functions. They might act through a variety of mechanisms, including chromatin modification, transcriptional and post-transcriptional regulation of gene expression, RNA splicing, and protein translation and turnover (Nie et al., 2012; Gutschner and Diederichs, 2012) and interaction with protein and microRNAs (Ma et al., 2012) (Figure 9). As a consequence, deregulation of lncRNAs can play a significant role in carcinogenesis (Fang et al., 2014; Gupta et al., 2010; Garding et al., 2013). Here we list a number of cancer-associated lncRNAs, which are often aberrantly expressed and actively implicated in various tumoral processes in human cancer (Table 3).

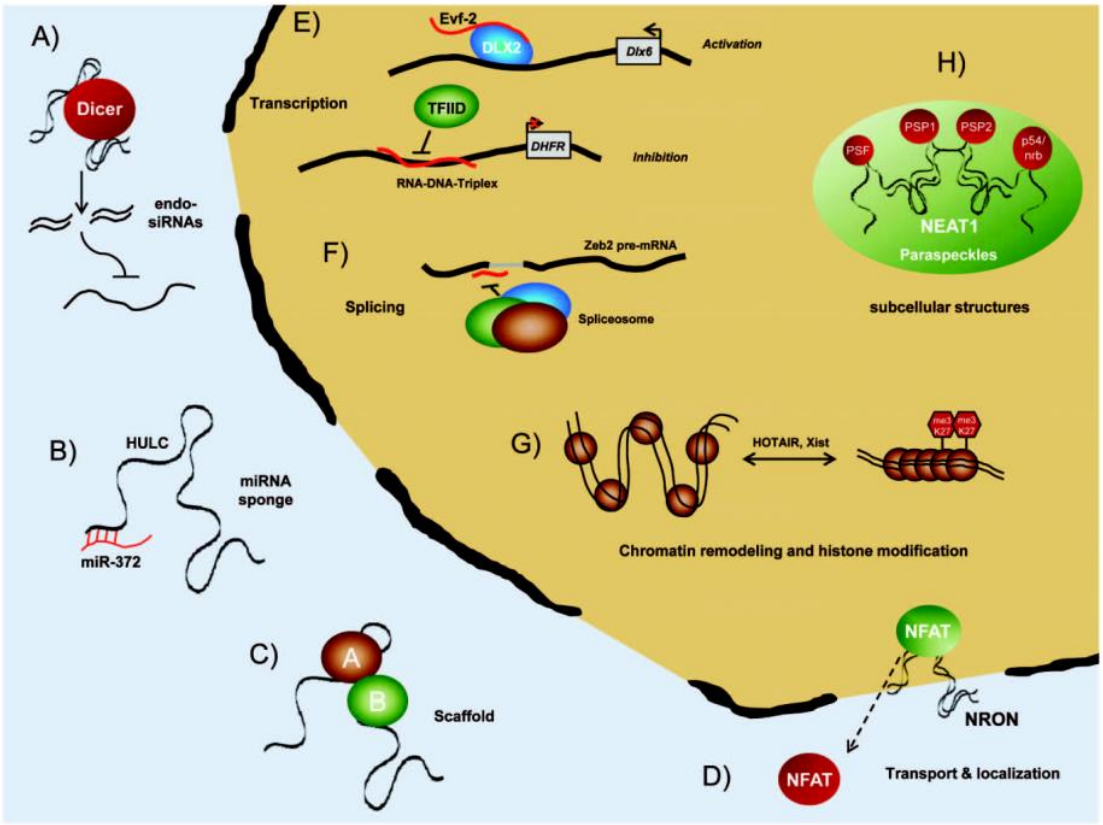

Figure 9. Graphical display of mechanisms by which lncRNAs function in cells (Gutschner and Diederichs, 2012). LncRNAs can function in a variety of ways. Overall, lncRNAs are able to alter expression of target genes, affect protein localization and activity (D) and play an important role in the formation of cellular substructures (such as paraspeckles) and protein complexes (such as scaffold) (C;H) (Clemson et al., 2009). (A) LncRNAs can be degraded into small endo-siRNAs, which are capable of silencing target gene expression. (B) LncRNAs function as "miRNA sponges", which inactivate target miRNAs expression and alter the expression of downstream genes of these miRNAs (Wang et al., 2010). (D) LncRNAs may function via interaction with proteins, for instance, NRON (non-coding repressor of NFAT) can bind to the transcription factor NFAT (nuclear factor of activated T cells) and transport NFAT from nuclear to cytoplasm, which suppresses NFAT target gene expression (Willingham et al., 2005). (E) Moreover, lncRNA may either recruit or block transcription factors to bind to target gene promoters, which leads to activation or degradation of target gene transcription (Feng et al., 2006; Martianov et al., 2007). (F) LncRNAs can modulate alternative splicing of target mRNAs via formation of the spliceosome complex (Beltran et al., 2008). (G) LncRNAs may also participate in the epigenetic regulation, they can regulate chromatin status via interaction with chromatin remodeling complexes or histone modification (Rinn et al., 2007; Zhao et al., 2008).

3.2 LncRNAs and proliferation

One important feature that cancer possesses is unlimited growth without the stimulation of external factors. Normal cells are able to produce proliferation promoting or inhibiting factors which tightly control the number of cells and functions, however, malignant tumor cells are able to escape from proliferation signals and obtain uncontrolled growth through a wide range of ways, such as hypoxia, dysregulation of cell cycle genes such as the Rb pathway (INK4-cyclin D-cdk4/6-Rb) and Cyclins D and E as well as activation of signaling pathways such as Wnt/β-catenin signaling, PI3K/Akt/mTOR signaling, Notch signaling and NF-κB signaling (Feitelson et al., 2015). In the past ten years, there was increasing evidence demonstrating lncRNAs affect the proliferation of cancer cells. Sun et al. (Sun et al., 2015) used GRO-seq and RNA-seq to annotate lncRNAs in MCF-7 breast cancer cell line and found about 1900 lncRNAs, more than 700 of which are newly identified lncRNAs. LncRNA152 and lncRNA67 were functionally characterized further in breast cancer, these two lncRNAs are

upregulated in breast tumors. Silencing their expression by siRNA-mediated deletion greatly inhibited cellular proliferation in MCF-7 and T47D breast cancer cell-lines. In contrast, enhanced expression of LncRNA152 and lncRNA67 in part rescued the growth inhibition by siRNA knockdown in MCF-7 cells. In addition, LncRNA152 and lncRNA67 are implicated in the regulation of cell cycle and estrogen receptor pathway. Knockdown of either lncRNA increased the number of cells in G1 phase and reduced the fraction of cells in S phase. Most importantly, lncRNA152 and lncRNA67 interacted with estrogen signaling pathway, which might in part account for their control of cell cycle. Sun et al (Sun et al., 2015) found that estrogen affected the expression of lncRNA152 and lncRNA67, with lncRNA152 upregulated and lncRNA67 downregulated. Estrogen treatment in part reduced the inhibitory effect on cellular growth of MCF-7 by knockdown of lncRNA152; however, silencing of either lncRNA repressed the expression of many estrogen-regulated target genes. Another evidence of lncRNAs playing a role in cancer proliferation is PCAT-1 (prostate cancer associated transcript 1). PCAT-1 is overexpressed in high-grade and metastatic prostate cancer samples. Knockdown and enhanced expression of PACT-1 led to decreased proliferation rate and modest increase in cellular growth, respectively. In addition, downregulation of PCAT-1 by siRNA-mediated knockdown caused deregulation of 370 protein-coding genes, among which 255 are upregulated and 115 downregulated. Gene ontology enrichment analyses found that upregulated genes were related to cell cycle and mitosis, suggesting that PCAT-1 might contribute to proliferation through transcriptional regulation of cell cycle and mitosis-associated genes in prostate cancer (Prensner and Chinnaiyan, 2011).

An alternative mechanism sustaining proliferation involves cancer cells that are able to escape proliferation suppression operated by tumor suppressor genes, such as TP53, PTEN and RB. External or internal stimuli, such as radiation and hypoxia activate these tumor suppressor genes, leading to cell cycle disruption or apoptosis. Recent studies have shown that lncRNAs are involved in the inhibition of tumor suppressor genes in diverse ways. H19, located on chromosome 11p15.5, is markedly increased in gastric cancer cell lines and cancer samples. Enhanced H19 expression decreases P53 activity and protein levels of the p53 target Bax, leading to promotion of cell proliferation and reduction of cell apoptosis (F. Yang et al., 2012).

Expression of Alu-mediated p21 transcriptional regulator (APTR) is negatively correlated to

that of p21 in gliomas. APTR inhibits the transcription of CDKN1A/p21 via recruitment of the PRC2 complex to the promoter of CDKN1A/p21, leading to activation of cell proliferation in HCT116 and to G1-S arrest in MCF10A cancer cells. The localization of APTR to the p21 promoter is mediated by the Alu (c-Alu) element embedded in APTR. Expression of p21 is induced and expression of APTR is reduced irrespective of p53 activity in human glioma cells, in response to cell stresses, such as heat shock and doxorubicin. This body of evidence supports that APTR represses p21 epigenetically via recruiting PRC2 to the p21 promoter (Negishi et al., 2014).

3.3 LncRNAs and invasion and metastasis

Cancer cells are able to invade and metastasize to form secondary tumors, which makes treatment of cancer highly challenging and causes high mortality rate. In order to successfully invade into healthy tissues, cancer cells have to go through multiple processes, including morphological changes, transition through lymphatic system and blood vessels and formation of micrometastases, eventually formation of a secondary tumor (Gutschner and Diederichs, 2012). Epithelial mesenchymal transition (EMT) is a developmental regulatory process which plays a great role in the regulation of cancer invasion and metastasis (Yilmaz and Christofori, 2009; Polyak and Weinberg, 2009). During EMT, epithelial cells that are non-mobile, polarized, embedded via cell-cell junctions are transformed into invasive mesenchymal cells that are individual, non-polarized and mobile. Several important factors are critical to the EMT process, such as E-cadherin (CDH1) and N-Cadherin (CDH2). As a critical cell-to-cell adhesion molecule, E-cadherin is frequently downregulated or inactivated in human cancers (Berx and van Roy, 2009; Cavallaro and Christofori, 2004). Upregulation of E-cadherin therefore represses cancer invasion and metastasis, E-cadherin is under strict control by multiple factors, such as Snail1 (Snail), Snail2 (Slug), ZEB1 (δEF1), ZEB2 (Sip1), E47, and Twist which are transcriptional repressor of E-cadherin (Peinado et al., 2007) and receptor tyrosine kinase or Src which mediates phosphorylation and degradation of E-cadherin (Beltran et al., 2008; Yilmaz and Christofori, 2009). N-Cadherin (CDH2), that is normally expressed in nervous tissues and mesenchymal cells, forms homophilic cell-cell

adhesion junctions. Its expression can be upregulated by collagen I, α2β1-integrin and Twist (Alexander et al., 2006; Shintani et al., 2008).

An increasing number of evidences show lncRNAs are implicated in cancer invasion and metastasis in a variety of ways. The exemplary lncRNA MALAT1 (Metastasis-Associated Lung Adenocarcinoma Transcript 1, MALAT-1) shows abundant expression in diverse cell types and high conservation across various species (Gutschner et al., 2011; Tripathi et al., 2010). MALAT1 is upregulated in several cancer types including lung cancer, uterine endometrial stromal sarcoma and hepatocellular carcinoma (Ji et al., 2003; Guo et al., 2010; Lin et al., 2007; Tano et al., 2010). MALAT1 plays an active role in cancer metastatic process, for instance, it regulates motility-associated genes and enhances cellular motility of lung cancer cells, depletion of MALAT1 by siRNAs reduces the expression of CTHRC1, CCT4, HMMR or ROD1, which impairs cell motility in lung adenocarcinoma (Tano et al., 2010). Nude mice with depletion of MALAT1 expression developed less number of lung tumor nodules and metastases (Schmidt et al., 2011; Gutschner et al., 2013). Moreover, MALAT1 also promotes cellular proliferation and metastasis of cervical cancer cells, silencing MALAT1 expression results in deregulation of apoptosis pathway related genes, such as caspase-8, caspase-3, Bcl-2 and Bcl-xL in cervical cancer (Guo et al., 2010). MALAT1 is involved in the regulation of epithelial-mesenchymal transition (EMT) associated genes, Downregulation of MALAT1 expression leads to downregulation of ZEB1, ZEB2 and Slug and upregulation of E-cadherin in bladder cancer, which induces epithelial-to-mesenchymal transition and metastasis in bladder cancer (Ying et al., 2012).

Another cancer metastasis-associated lncRNA is HOTAIR (HOX Antisense Intergenic RNA), HOTAIR expression is upregulated in primary and metastatic tumors of different cancer types, including breast cancer (Gupta et al., 2010), colorectal cancer (Kogo et al., 2011), pancreatic cancer (Kim et al., 2013), hepatocellular carcinoma (Geng et al., 2011), gastrointestinal stromal cancer (Niinuma et al., 2012) and oesophageal squamous cell carcinoma (X. Li et al., 2013). HOTAIR expression is high in breast cancer that are predisposed to metastasize, and its inhibition blocks metastasis in mouse models (Gupta et al., 2010). HOTAIR plays an important role in epigenetic regulation, enhanced expression of HOTAIR interacts with PRC2 (polycomb repressive complex 2) to alter H3K27 methylation , leading to changes of target

gene expression in epithelial breast cancer cell and increased cancer metastasis, in contrast, knockdown of HOTAIR suppresses cancer invasion and metastasis (Gupta et al., 2010). HOTAIR expression is upregulated in hepatocellular carcinoma compared to adjacent normal tissues, increased expression of HOTAIR indicates recurrent HCC and poor survival (Yang et al., 2011), furthermore, HOTAIR might serve as a potential indicator of lymph node metastasis in liver cancer; downregulation of HOTAIR expression greatly leads to decreased cellular metastasis and viability in liver cancer cells (Geng et al., 2011).

The third metastasis-involved lncRNA is H19, upregulation of H19 expression is observed in hepatocellular carcinoma (Matouk et al., 2007), bladder cancer (Luo et al., 2013) and lung cancer (Matouk et al., 2014). H19 has been demonstrated to actively contribute to tumoral metastasis and invasion through multiple mechanisms. H19 directly affects the expression of the key players of EMT process, H19 expression is negatively correlated with E-cadherin and assists in binding of Ezh2, an epigenetic regulator, to the promoter of E-cadherin and indirectly activates Wnt-βcatenin, which leads to transcriptional repression of E-cadherin in bladder cancer (Luo et al., 2013). Moreover, H19 suppresses E-cadherin expression through a positive feedback loop between Slug and H19/miR-675, in which H19 induces Slug expression through miR-675-implicated mechanism, and upregulation of Slug further activates H19 promoter and enhances H19 expression levels in lung cancer (Matouk et al., 2014). H19 is also shown to regulate tumor metastasis via epigenetic activation of miR-200 family in liver cancer, ectopic expression of H19 interacts with the HnRNPU/PCAF/RNA PolII complex and enables the binding of the complex to the promoter of miR-200 family, which activates miR-200 family via enhancing histone H3 acetylation, thus H19 can epigenetically activate the miR-200 pathway, leading to induction of mesenchymal-to-epithelial transition and the inhibition of cancer metastasis (L. Zhang et al., 2013).

3.4 LncRNAs and apoptosis

Apoptosis plays an important role in a wide range of diseases, including cancer. Cells initiate apoptotic processes in response to external stimuli, such as glucocorticoids, radiation, hypoxia and infection. Apoptotic processes are executed by two main mechanisms, including the

extrinsic death receptor pathway and the intrinsic mitochondrial apoptosis pathway. The extrinsic pathway mainly consists of three parts: the death ligands, such as tumor necrosis factor and Fas ligand, transmembrane receptors, such as the type I TNF receptor and Fas receptor as well as adaptor proteins, such as Fas-associated death domain and TNF receptor-associated death domain. Death ligands bind to the extracellular domain of transmembrane receptors, and the death receptors interact with adaptor proteins, which leads to the formation of a death-inducing signaling complex (DISC) between Pro-caspase-8 and adaptor proteins and activation of Pro-caspase-8 (Khosravi-Far and Esposti, 2004). The mitochondrial apoptosis can be achieved in many ways. DNA damage initiates apoptosis through activating the tumor-suppressor protein p53, which consequentially upregulates the expression of pro-apoptotic genes such as DR-5, BAX, BAK, NOXA, PUMA and downregulates the expression of anti-apoptotic genes such as Bcl-2 and survivin (Goldar et al., 2015). Moreover, intracellular stimuli can affect the permeability of mitochondrial membrane, initiate mitochondrial swelling via the BCL-2 family which includes 25 pro- and anti-apoptotic members (Chipuk et al., 2004). The imbalance among these pro-apoptotic and anti-apoptotic Bcl-2 family members increases the permeabilization of mitochondrial membranes and leads to leakage of cytochrome C and other mitochondrial proteins. For instance, the release of mitochondrial proteins such as SMACs (second mitochondria-derived activator of caspases) deactivates inhibitor of apoptosis proteins (IAPs) and indirectly promotes the activities of caspases; another apoptotic protein , cytochrome c, is released by mitochondria through the formation of the mitochondrial apoptosis-induced channel (MAC). Cytochrome C together with apoptotic protease activating factor-1 and ATP form a complex “apoptosome”, which transforms pro-caspase-9 into its active form of caspase-9, activates caspase-3 and eventually results in cell death (Zou et al., 1997; Jin et al., 2005).

A number of lncRNAs has been observed to affect cancer apoptosis pathways, such as PCGEM1, CUDR and PANDAR. PCGEM1 is overexpressed and shows anti-apoptotic effect in prostate cancer (Srikantan et al., 2000). Overexpression of PCGEM1 led to expression delay of p53 and p21 and remarkably decreased cleaved caspase 7 and PARP expression in doxorubicin-treated LNCaP cells.The apoptotic inhibition is highly androgen-dependent, as mutations of androgen could diminish this effect (Liebert and Gene, 2006). Another anti-

apoptotic lncRNA is CUDR (cancer upregulated drug resistant) displaying inhibitory effect on drug-induced apoptosis, such as doxorubicin and etoposide in squamous carcinoma cells A431. Enhanced expression of CUDR downregulates the effector caspase 3, which might account for this inhibitory function of apoptosis (Jin et al., 2005; Khosravi-Far and Esposti, 2004).

However, many lncRNAs play a pro-apoptotic role in cancer, such as PANDAR (Han et al., 2015), INXS (DeOcesano-Pereira et al., 2014) and GAS5 (Kino et al., 2010). PANDAR is lowly expressed in non-small cell lung carcinoma (NSCLC), and downregulation of PANDAR expression correlates negatively with great tumor size and late tumor stage. Enhanced expression of PANDAR could greatly increase the apoptosis rate of lung cancer cell lines, A549 and SPC-A1, the apoptosis-inducing effect is in part rescued by upregulation of P53. Overexpression of PANDAR could induce the expression of pro-apoptotic proteins (Bax and Bad) and inhibit anti-apoptotic protein (Bcl-2), which leads to the activation of caspase-3 and induction of apoptosis in NSCLC cells (Han et al., 2015).

INXS is a 1903 nts pro-apoptotic lncRNA that is transcribed from the opposite strand of the BCL-X genomic locus, INXS is significantly less abundant in kidney cancer in comparison with adjacent normal tissues. Treatment of apoptosis-inducing agents, UV-C light exposure and anti-cancer agent sulforaphane (SFN), led to increased expression of INXS and activation of caspases 3, 7 and 9 in 786-O kidney tumor cells, siRNA-mediated deletion of INXS could greatly diminish such an effect. Overexpression of INXS resulted in a pronounced accumulation of pro-apoptotic BCL-XS and activated activation of caspases 3, 7 and 9 as well as a decrease of anti-apoptotic BCL-XL abundance, thus inducing apoptosis in 786-O cells. Furthermore, tumor weight was reduced by increased BCL-XS expression after injection of INXS-expressing plasmid in mouse xenograft model. All these evidences support that INXS is an apoptosis-inducer in kidney cancer (DeOcesano-Pereira et al., 2014).

GAS5 (Growth Arrest-Specific 5), firstly identified in mouse NIH3T3 fibroblasts, is downregulated in various cancer types, such as leukemia and breast cancer (Coccia et al., 1992; Schneider et al., 1988). GAS5 interacts with DNA binding domain of the glucocorticoid receptors and blocks the DNA glucocorticoid response elements to bind these receptors, which inhibits the glucocorticoid-mediated transcription of anti-apoptotic genes like cellular

inhibitor of apoptosis 2 (cIAP2) and leads to cellular apoptosis (Kino et al., 2010).

3.5 LncRNAs and cell cycle

The cell division cycle consists of quiescent/senescent (G0) phase, Interphase (G1, S and G2 phase) and Cell division (M) phase. The G0 phase is a resting phase in which cells have finished division. Interphase is the stage where cells prepare for mitosis, including the G1 phase which supplies proteins and increases the number of organelles. The S phase is that for DNA synthesis, and the G2 phase is that for cell growth. Lastly, cell growth stops and cells are divided into two daughter cells in the M phase. The cell cycle is under strict regulation of cyclin-dependent kinases (CDKs) and their related pathways in mammalian cells. The CDKs bind to cyclins, including cyclins A, B, D, and E, and form CDK-cyclin complexes which phosphorylate and activate their target genes, enabling cell cycle progression (Morgan, 1995). For instance, in response to extracellular signals, such as growth factors, Cyclin D binds to CDK4 and forms the cyclin D-CDK4 complex which in turn phosphorylates the retinoblastoma susceptibility protein (Rb) and its family members, p107 and p130 and activates E2F transcription in the late G1 phase. The activation of E2F leads to activation of multiple growth-promoting genes such as cyclin E, DNA polymerase (Weinberg, 1995; Kitagawa et al., 1996). Cyclin E-CDK2 phosphorylates pRB as well as several proteins involved in DNA replication to push the cell from G1 to S phase (Hwang and Clurman, 2005). Moreover, the cell cycle is negatively regulated by CDK inhibitors, such as p15, p16, p18, p21, p27, and p57 which inhibit the activities of cyclin–CDK complexes through specific binding to their targets (Sherr and Roberts, 1999; Vidal and Koff, 2000).

A number of lncRNAs plays important roles in the progression of the cancer cell cycle through regulation of expression of critical cell cycle genes, such as Purα, CDKs and cyclins (Bida et al., 2015; Liu et al., 2012; Tripathi et al., 2013). MA-linc1 (Mitosis-Associated Long Intergenic Non-Coding RNA 1) locates on the chromosome 5 and consists of three exons, it functions as a transcriptional target gene of E2F1. Knockdown of MA-linc1 alters cell cycle distribution of the human osteosarcoma cell line U2OS, characterized by a reduction of G1 phase cells and an increase in cancer cells at G2/M and S phase. Moreover, silencing expression of MA-linc1 led to decreased mitosis exit in M phase-arrested cells. The

mechanism underlying the cell cycle regulation of MA-linc1 can be partly mediated by cis repression of the expression of its neighboring gene Purα (DeOcesano-Pereira et al., 2014), which is often deleted in cancers and whose aberrant expression arrests cell cycle progression (Bida et al., 2015; Gallia et al., 2000). In support of the above findings, knockdown of MA-linc1 induces cellular apoptosis initiated by the antimitotic drug, Paclitaxel and deletion of Purα could rescue such an enhancement of apoptosis (Bida et al., 2015).

The gadd7 (growth-arrested DNA damage-inducible gene 7) lncRNA (DeOcesano-Pereira et al., 2014) is another important lncRNA that controls cell-cycle progression. It was firstly identified from Chinese hamster ovary (CHO) cells owing to its abundant expression after UV irradiation (Hollander et al., 1996). Depletion of gadd7 leads to an increase of cellular proliferation and cell cycle redistribution, with a remarkable reduction of G1 phase cells and an accumulation of G2/M and S phase cells in response to DNA damage caused by UV radiation, suggesting that gadd7 may affect G1/S transition. Following UV radiation, gadd7 expression is induced and it directly binds to TAR DNA-binding protein (TDP-43) and dissociates TDP-43 from cyclin-dependent kinase 6 (Cdk6) mRNA, which leads to Cdk6 mRNA decay and the regulation of G1/S checkpoint (Liu et al., 2012).

P53, as a tight regulator of the cell cycle, is able to control both G1 and G2/M checkpoints (Schwartz and Rotter, 1998). Many lncRNAs function as cell cycle regulators via P53-mediated cell cycle control (Léveillé et al., 2015;Sánchez et al., 2014) , such as PR-lncRNA-1, PR-lncRNA-10 and RoR. PR-lncRNA-1 and PR-lncRNA-10, localized in the nucleus of cells, are two transcriptional targets of P53. Gene expression analysis revealed that PR-lncRNA-1and PR-lncRNA-10 depletion led to dysregulation of several genes associated to cell cycle control and apoptosis, which are p53 downstream target genes. Moreover, PR-lncRNA-1 and PR-lncRNA-10 are essential to the binding of p53 to p53 target genes, such as SERPINB5, CDKN1A, BCL2L1and BBC3 genes. Silencing the expression of PR-lncRNA-1and PR-lncRNA-10 caused a significant increase of cell proliferation, and decrease of cell apoptosis. Deletion of PR-lncRNA-1 and PR-lncRNA-10 increased the number of cells in S-phase of cell cycle in HCT116 cells. Overall, these findings support that PR-lncRNA-1 and PR-lncRNA-10 contribute to an induction of apoptosis and cell cycle arrest via the p53 signaling pathway (Ji et al., 2003; Guo et al., 2010; Lin et al., 2007; Tano et al., 2010; Sánchez et al.,

2014). Another lncRNA named RoR interacts with the heterogeneous nuclear ribonucleoprotein I (hnRNP I) through binding of hnRNP I to a 28-base RoR motifs, which enables to suppress the expression of p53 in response to ultraviolet C (UVC). As a result, RoR reduced the p53-mediated apoptosis in MCF-7 cells and G2/M arrest in HCT-116 WT cells (A. Zhang et al., 2013).

Table 3. A list of experimentally characterized cancer-related lncRNAs

	LncRNA	**Expression**	**Cancer type**	**Function**	**Reference**
Proliferation	LncRNA152 lncRNA67	Up-regulated	breast cancer	growth-promoting	(Sun et al., 2015)
	PACT-1	Up-regulated	Prostate cancer	growth-promoting	(Prensner et al., 2011)
	APTR		Colon cancer, glioblastoma	growth-promoting	(Negishi et al., 2014)
	H19	Up-regulated	Hepatocellular, bladder, lung cancer, breast and gastric cancer	growth-promoting, metatasis inducer	(Matouk et al., 2007; Barsyte-Lovejoy, 2006; Berteaux et al., 2005; F. Yang et al., 2012; Luo et al., 2013; Matouk et al., 2014)
	Sox2ot	Up-regulated	Lung squamous cell carcinomas (SCCs)	growth-promoting	(Hou et al., 2014)
	GAS5	Down-regulated	Leukemia,non-small-cell lung cancer,bladder cancer	growth-inhibiting, apoptosis inducer	(Braconi et al., 2010; Coccia et al., 1992; Shi et al., 2013; Z. Liu et al., 2013)
	HULC	Up-regulated	Liver,gastric cancer	miR-372 sponge, growth-promoting, metasasis inducer and apoptosis inhibitor	(Wang et al., 2010; Zhao et al., 2014)
	PCNA-AS1	Up-regulated	Hepatocellular carcinoma	growth-promoting	(Yuan et al.,2014)
	PRNCR1	Up-regulated	Prostate cancer	growth-promoting	(Chung et al., 2011)
	ANRIL	Up-regulated	Prostate cancer, acute lymphoblastic leukemia, glioma, melanoma	growth-promoting	(Yap et al., 2010;Cunnington et al., 2010; Iacobucci et al., 2011)
	T-UCR uc.338	Up-regulated	Liver cancer	growth-promoting	(Braconi et al., 2010)
	SPRY4-IT1	Up-regulated	Melanoma	growth-promoting, apoptosis inhibitor	(Khaitan et al., 2011)
	PlncRNA-1	Up-regulated	Esophageal squamous carcinoma	growth-promoting	(Wang et al., 2014)
	HNF1A-AS1	Up-regulated	Oesophageal adenocarcinoma	growth-promoting ,metasasis inducer	(X. Yang et al., 2014)

	ncRAN	Up-regulated	Bladder cancer	growth-promoting ,metasasis inducer	(Zhu et al., 2011)
	GHET1	Up-regulated	Gastric and bladder cancer	growth-promoting	(F. Yang et al., 2014; Li et al., 2014)
	LOC285194 BC040587	Down-regulated	Osteosarcoma,colon cancer	growth-inhibiting	(Q. Liu et al., 2013; Pasic et al., 2010)
	PTENP1		PTENP1 locus is selectively lost in human cancer	growth-inhibiting	(Poliseno et al., 2010)
	MEG3	Down-regulated	Brain cancer, non-small cell lung cancer	growth-inhibiting	(Zhang et al., 2003;Lu et al., 2013)
	HOTTIP	Up-regulated	Pancreatic cancer	growth-promoting, metasasis inducer, apoptosis inhibitor	(Cheng et al., 2015)
	PCAN-R1 PCAN-R2	Up-regulated	Prostate cancer	growth-promoting	(Du et al., 2013)
Metastasis	ARLTS1	Down-regulated	Lung cancer	growth-inhibiting	(Yendamuri et al., 2007)
	MALAT1	Up-regulated	lung cancer, uterine endometrial stromal sarcoma, cervical cancer and hepatocellular carcinoma	Metatasis inducer	(Ji et al., 2003; Guo et al., 2010; Lin et al., 2007; Tano et al., 2010).
	HOTAIR	Up-regulated	Breast cancer,liver cancer	Metastasis inducer	(Gupta et al., 2010; Geng et al., 2011)
	BANCR	Up-regulated	Melanoma	Metatasis inducer	(Flockhart et al., 2012)
	UCA1	Up-regulated	Tongue squamous cell carcinoma	Metasasis inducer	(Fang et al., 2014)
	lncRNA-EBIC	Up-regulated	Cervical cancer	Metasasis inducer	(N. Sun et al., 2014)
	AOC4P	Down-regulated	Hepatocellular carcinoma	Metasasis inhibitor	(Wang et al., 2015)
	ZEB1-AS1	Up-regulated	Hepatocellular carcinoma	Metasasis inducer	(Li et al., 2015)
	lnc-ATB	Up-regulated	Breast cancer	Metasasis inducer	(Shi et al., 2015)
	HNF1A-AS1	Up-regulated	Lung cancer	Metasasis inducer	(Wu et al., 2015)
	DRAIC/PCAT29	Down-regulated	Prostate cancer	Metasasis inhibitor	(Sakurai et al., 2015)
	HOTTIP and HOXA13	Up-regulated	Hepatocellular carcinoma	Metasasis inducer	(Quagliata et al., 2014)
	treRNA	Up-regulated	Breast cancer	Metasasis inducer	(Gumireddy et al., 2013)
	ESCCAL-1	Up-regulated	Esophageal squamous cell carcinoma (ESCC)	Metasasis inducer, apoptosis inhibitor	(Hao et al., 2015)
	NKILA	Down-regulated	Breast cancer	Metasasis inhibitor	(Liu et al., 2015)
Apoptosis	PCGEM1	Up-regulated	Prostate cancer	Apoptosis inhibitor, growth-promoting	(Petrovics et al., 2004)
	CUDR	Up-regulated	Human squamous cancer	Apoptosis inhibitor	(Jin et al., 2005;Khosravi-Far

					and Esposti, 2004).
	PANDAR	Down-regulated	non-small cell lung carcinoma (NSCLC)	Apoptosis inducer	(Han et al., 2015)
	INXS	Down-regulated	Kidney cancer	Apoptosis inducer	(DeOcesano-Pereira et al., 2014)
	TUG1	Up-regulated	Hepatocellular carcinoma (HCC)	Apoptosis inducer, growth-promoting	(M. Huang et al., 2015)
	uc.73a	Up-regulated	Leukemia, colorectal cancer	Apoptosis inducer	(Calin et al., 2007)
	uc002mbe.2		Liver cancer	Apoptosis inducer	(H. Yang et al., 2013)
	LincRNA-p21		Lung cancer, sarcoma, lymphoma	Apoptosis inducer	(Huarte et al., 2010)(
Cell cycle	AK126698	Down-regulated	Non-small-cell lung cancer	Apoptosis inducer	(Y. Yang et al., 2013)
	MA-linc1		osteosarcoma	Cell cycle G1 phase arrest, apoptosis inducer	((Bida et al., 2015)
	gadd7		CHO-K1 cells (Hamster Chinese ovary)	G1/S checkpoint, growth-inhibiting	(Liu et al., 2012)
	PR-lncRNA-1 and PR-lncRNA-10	Down-regulated	Colorectal cancer	Cell cycle G1 phase arrest, apoptosis inducer, growth-inhibiting	(Sánchez et al., 2014)
	lincRNA-RoR (RoR)		Breast cancer, colon cancer	Inhibition of G2/M arrest, apoptosis inhibitor,	(A. Zhang et al., 2013)
	Linc00152	Up-regulated	Gastric cancer	Cell cycle G1 phase arrest, growth-promoting , apoptosis inhibitor	(Zhao et al., 2015)
	lncRNA-HEIH	Up-regulated	Hepatocellular carcinoma (HCC)	G0/G1cell cycle arrest	(Yang et al., 2011)
Others	DD3(PCA3)	Up-regulated	Prostate canccer	A diagnostic marker	(Kok et al., 2002)
	XIST	Lost in female breast, ovarian, and cervical cancer cell lines	Breast, ovarian, and cervical cancer	X chromosome silencing	(McHugh et al., 2015)

3.6 Development of computational tools for functional lncRNA prediction

Through gene regulation or other mechanisms, lncRNAs are emerging as important players in the cancer paradigm, acting as proto-oncogenes, tumor suppressor genes and drivers of metastatic transformation. Even though an increasing number of lncRNAs have been functionally characterized, the biological functions of the majority of lncRNAs remain unknown. Therefore, bioinformatics tools are urgently needed to prioritize cancer-related lncRNAs. Currently, more and more studies are being developed to explore methods to identify either cancer or disease-related lncRNAs. Table4 summarizes the computational

approaches used to predict functional lncRNAs.

3.6.1 Recurrent Somatic Copy-number Alteration-based Approach

Du et al. selected lncRNAs in recurrent somatic copy-number alterations (SCNAs) (gain) regions as candidate drivers, such as PCAN-R1 or PCAN-R2 which are the two most significantly differentially expressed lncRNAs between tumor and normal prostate tissues. Knockdown of them resulted in substantial decrease in both cell growth and colony formation in the androgen-dependent prostate cancer cell line LNCaP, suggesting they have tumor-promoting functions in prostate cancer (Du et al., 2013).

3.6.2 Coexpression with Coding Genes Approach

Guttman et al. have developed a coexpression based method to functionally characterize lncRNAs. They ranked protein coding genes according to their correlation coefficients of expression levels with each lncRNA, and then performed a Gene Set Enrichment Analysis (GSEA) on high ranking genes to identify function enrichment for each lncRNA. Application of this coexpression method to 1,600 lncRNAs found that lncRNAs are actively implicated in a wide range of functional processes, including cell proliferation, development and embryonic stem cell pluripotency(Guttman et al., 2009).

Liao et al constructed a coding–non-coding gene co-expression (CNC) network which employs two different strategies to predict functions of lncRNAs, including the network hub-based method and network modules. The hub-based method determines lncRNA functions based on gene ontology (GO) enrichment analysis of surrounding protein coding genes. The authors use a Markov cluster algorithm (MCL) to search for coexpressed functional modules composing either non-coding or coding genes in the CNC network, and then assign functions to lncRNAs based on module functions. Application of the CNC method to 340 mouse lncRNAs found these lncRNAs have functions involving organ or tissue development, cellular transport, and metabolic processes (Liao et al., 2011). Liu et al developed a computational framework to prioritize disease-associated lncRNAs based on lncRNA, gene expression profile and gene-disease association data. They obtained expression profiles of 21626 lincRNAs generated by RNA-sequencing of 22 human tissues or cell types (Karolchik, 2004), 17080 genes from RNA sequencing of 73 human tissue or cell types (Su et al., 2004) and gene-disease associations from the DisGeNET database (Bauer-Mehren, 2010). They first

associated lncRNAs to tissue-specific diseases by combining high tissue specificity scores and high expression levels of lncRNAs in that tissue. Secondly, for non-tissue-specific lncRNAs, Spearman rank correlation coefficients were calculated between protein coding genes and each lncRNA to obtain a set of co-expressed genes. The hypergeometric distribution test for the set of genes co-expressed with each lncRNA was then used to predict potential lncRNA-associated diseases (Liu et al., 2014). Implementation of this computational framework enabled identification of 2272 potential lincRNA-associated diseases and novel lncRNAs for human diseases.

3.6.3 Network-based systems

Long non-coding RNA global function predictor ('lnc-GFP') integrates gene expression and protein interaction data to functionally annotate lncRNAs. The authors use a bi-colored network in which vertices represent protein-coding genes and lncRNAs, and edges stand for co-expression and protein interaction. lnc-GFP uses a global propagation algorithm in which 'function flow' from known function annotations for genes propagates on the network iteratively. The association score measuring how likely an unknown lncRNA can be functionally annotated combines the iterative propagation of the 'function flow' on the network and the previous knowledge score calculated between an unknown lncRNA and a given functional category (Guo et al., 2013). The authors claimed lnc-GFP is able to functionally characterize 94.9% of lncRNAs in their bi-colored network.

3.6.4 Interaction with Proteins and miRNAs Approach

Interaction of lncRNAs with proteins and miRNAs is a major path towards understanding the function of lncRNAs. Several methods have been developed to explore interactive properties of lncRNAs with proteins and miRNAs and indirectly predict their functions. Bellucci et al have developed catRAPID to assess the interaction propensities of lncRNAs with proteins using their physicochemical properties, including secondary structure, hydrogen bonding and van der Waals. The catRAPID method was trained on 592 protein-RNA pairs from the Protein Data Bank (Bellucci et al., 2011). catRAPID has a prediction accuracy of 0.89, which is validated with experimentally supported protein associations annotated in the NPInter dataset (Wu, 2006). Jeggari et al have developed the program miRcode which aims to predict putative target sites of microRNAs in 10,419 lncRNAs. The miRcode program is constructed mainly based on two criteria, complementarity to seed regions, the 2rd-8th bases from the 5′-end of

the microRNA, and evolutionary conservation, as assessed from 46 vertebrate genome alignments (Jeggari et al., 2012).

Table 4. Summary of computational approaches for predicting disease or cancer related functional lncRNAs

Name	Based on	Cancer-specific	References
Recurrent SCNAs -based Approach	Recurrent somatic copy-number alterations (SCNAs) and differential expression of lncRNAs	Yes	(Du et al., 2013)
Guttman et al 's coexpression based method	Coexpression with coding genes and Gene Set Enrichment Analysis (GSEA)	No	(Guttman et al.,2009)
a CNC network	Coexpression with coding genes, gene ontology (GO) enrichment analysis and	No	(Liao et al., 2011)
Liu et al's coexpression based method	Coexpression with coding genes and gene-disease associations	No	(Liu et al., 2014)
Zhao et al 's co-expression network	Coding-noncoding gene co-expression network	Yes	(Zhao,2014)
Hao et al 's co-expression network	Coding-noncoding gene co-expression network and differential expression	Yes	(Hao et al., 2015;. Hao,2015)
lnc-GFP	Gene expression and protein interaction and a global propagation algorithm	No	(Guo,2013)
catRAPID	RNA and protein interaction	No	(Bellucci,2011)
miRcode	Complementarity to seed regions and evolutionary conservation	No	(Jeggari,2012)

Even though much has been done to predict functional lncRNAs based on different algorithms, the computational prediction of lncRNA function is still in its infancy. Current methods mainly rely on the coexpression or interactive relation of lncRNAs with other molecules, such as protein coding genes, miRNAs, and proteins. However, they do not take into account the importance of cancer mutations to the formation of lncRNA functions.

Recently, Gonzalez-Perez' et al developed a novel approach, Oncodrive-fm, to identify cancer driver candidates. The rationale of Oncodrive-fm is cancer drivers tend to accumulate somatic mutations with high functional impact and any bias towards enrichment of variants with high functional impact indicates positive selection for the driver genes in the tumor. Oncodrive-fm (Gonzalez-Perez and Lopez-Bigas, 2012) applies SIFT, Polyphen2 and MutationAssessor to score the functional impact (FI) of each coding mutation, and calculates the average FI scores for the variants observed in each gene across all cancer samples. Cancer drivers display a shift toward accumulation of highly deleterious somatic mutations, therefore, they tend to have a high average FI score. For each gene and scoring system, Oncodrive-fm employs a

permutation test which randomly samples the same number of observed variants within the gene 1 million times and computes the average FI score for each sample, three P values are generated by comparing the average FI scores with a null distribution consisting of the 1million average FI scores. Application of Oncodrive-fm to 135 glioblastoma multiforme samples identified that most of recurrently mutated genes such as TP53, PTEN, NF1, PIK3R1, ERBB2, EGFR, RB1, PIK3CA, also show a high ranking function impact bias.

Chapter 4 – A Permutation-based model for lncRNA driver search

LI J, Drubay D, Michiels S, Gautheret D

Author contribution:

Jia LI was the main contributor to this study, he performed the whole experiment under the supervision of Professor Daniel Gautheret. Drubay Damien and Michiels Stephan gave statistical support to the permutation-based model. Jia LI firstly wrote the manuscript, Daniel Gautheret gave his suggestion and comments and further revised the manuscript.

4.1 Introduction

In the light of the pioneering study by Gonzalez-Perez (Gonzalez-Perez and Lopez-Bigas, 2012), we hypothesized that cancer-associated lncRNAs would also display such a bias towards variants with functional impact. We implemented five different scoring systems to measure the function effect of non-coding variants: CADD, funSeq2, GWAVA, our SNP and SOM scores (Chapter 2). We applied a permutation- based model to prioritize cancer-associated lncRNAs. For each lncRNA, the permutation-based model randomly takes the same number of observed variants and calculates the average functional scores 1 million times to form a null distribution and produces a P value via comparing the observed functional score to the null distribution. To further validate our hypothesis and the permutation model, we implemented the permutation model on 61 cancer-related lncRNAs and 547 cancer genes using cancer mutation data of liver cancer, lung cancer, CLL and melanoma. We observed experimentally validated cancer driver genes showed significantly higher positive selection and FI bias than non-cancer genes. Applying our permutation test to lncRNAs using five different scoring systems enabled us to prioritize hundreds of cancer-related lncRNA candidates for further experimental validation. We found our candidates show enrichment for evolutionary conserved regions and disease-causing variants. Furthermore, overall our approach opens the way to the detection of cancer-driving lncRNAs and non-coding elements of genes on a genome wide scale.

4.2 Results

4.2.1 Validation of the permutation-based model on cancer genes and lncRNAs

We applied five different scoring systems to measure the function effect of non-coding variants: CADD, funSeq2, GWAVA, our SNP and SOM scores (Chapter 2). For each lncRNA and scoring system, the permutation-based model randomly takes the same number of observed variants and calculates the average functional scores 1 million times to form a null distribution, a raw P value was generated via comparing the observed functional score to the null distribution. The raw P values from five independent permutation tests were adjusted using False Discovery Rate (FDR) (Yekutieli and Benjamini, 1999). Finally, we use z transform (Whitlock, 2005) to combine five different P values to form an uniform P value. In

order to validate our permutation-based model, we applied it to 547 cancer-related protein-coding genes annotated in the COSMIC database and 61 cancer-related lncRNAs manually curated from recent publications (Table S20). Cancer-related protein-coding genes have significantly lower adjusted positive selection P values than total genes (P value < 0.05 in all cases, Wilcoxon rank sum test, Figure 10A, Figure S8A , S9A, S10A, Table S7). Similarly, the adjusted P values of cancer-related lncRNAs are significantly lower than those of total lncRNAs (P value <0.05 in all cases except for the CADD model in CLL, Wilcoxon rank sum test, Figure 10B, Figure S8B , S9B, S10B, Table S7).

We obtained the top 10 recurrently mutated genes (RMGs) for hepatocellular carcinoma, lung adenocarcinoma, Chronic lymphocytic leukaemia-small lymphocytic lymphoma and Malignant melanoma from the COSMIC database and analyzed their adjusted P values (Table 5, Table S8, S9, S10). If we consider for instance lung cancer, 40%, 100%, 60%, 80%, 80% and 100% of RMGs show statistically significant results (adjusted P value < 0.05) using the CADD, funSeq2 , GWAVA, SNP, SOM and combined model respectively. SETBP1 was positively selected by all six models with significant statistical evidence (adjusted P value < 0.05). EGFR, TP53, STK11, NF1, ZNF521 and GRIN2A had adjusted P values below 0.05 by any five models (Table 5). Next, we ranked the adjusted P values computed by each model and found 10%, 80%, 10%, 50%, 50% and 80% of RMGs have the first ranking in CADD, funSeq2 , GWAVA, SNP, SOM and combined models respectively. The P values of ZNF521 were ranked first by all but the SOM model. Three adjusted P values of funSeq2, SNP, SOM and combined models were ranked first for STK11, SETBP1, NF1 and SMARCA4. These results support the hypothesis that cancer-associated genes and lncRNAs display a bias towards accumulation of non-coding variants with high functional impact.

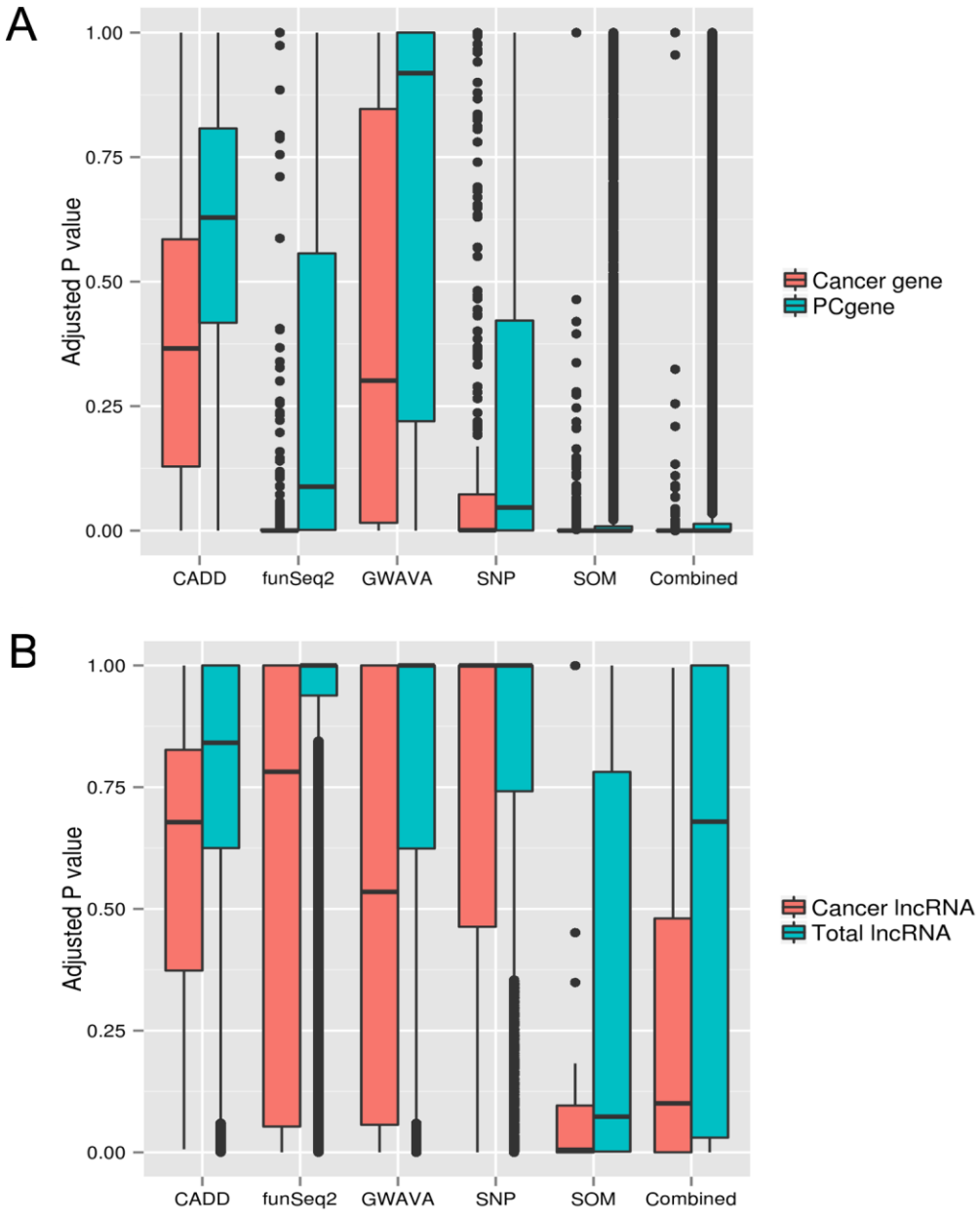

Figure 10. Distribution of adjusted P values for different gene classes. A. The comparison of adjusted P values computed by all permutation models between cancer-related genes and all genes; B. The comparison of adjusted P values computed by all permutation models between cancer-related lncRNAs and all lncRNAs.

Table 5. Adjusted P values and P value rankings of top 10 recurrently mutated genes in lung cancer

RMG	CADD	FunSeq2	GWAVA	SNP	SOM	Combined
	Adjusted Pvalue (Ranking of P value)					
EGFR	0,6691(3569)	0,0000(1)	0,0003(16)	0,0024(567)	0(1)	0(1)
TP53	0,0058(63)	0,0000(3)	0,7783(2741)	0,0095(1468)	0,0000(6)	0,0000(13)
KRAS	0,4988(2318)	0,0067(1018)	0,1030(1326)	0,1159(3819)	0,0577(3379)	0,0000(1351)
STK11	0,4653(2124)	0,0000(1)	0,0257(781)	0,0000(1)	0(1)	0(1)
SETBP1	0,0043(49)	0,0000(1)	0,0001(5)	0,0000(1)	0(1)	0(1)
SMARCA4	0,6627(3521)	0,0000(1)	0,5659(2400)	0,0000(1)	0(1)	0(1)
NF1	0,0225(183)	0,0000(1)	0,6368(2514)	0,0000(1)	0(1)	0(1)
CDKN2A	0,4503(2025)	0,0000(1)	0,0000(2)	1,0000(7507)	0,0017(824)	0(1)
ZNF521	0,0000(1)	0,0000(1)	0,0000(1)	0,0000(1)	0,4097(4769)	0(1)
GRIN2A	0,5151(2439)	0,0000(1)	0,0001(6)	0,0086(1383)	0,0024(994)	0(1)
Number of unique P values	6123	7508	3168	7507	5175	6891

4.2.2 General characteristics of driver candidates

We ran the permutation-based method to prioritize cancer-related PC genes and lncRNAs, using 1,613,031 non-coding variants from the same lung cancer data as in Chapter 2. We define driver candidates as PC genes and lncRNAs whose adjusted P values are less than 0.05. Overall, 180 to 10403 lncRNAs and 595 to 12797 PC genes meet the selection criteria, depending on the scoring system used (Table 6, Table S11, S12, S13). Overall, the CADD model detected fewer driver candidates and their size was longer compared to driver candidates identified by other models. In contrast, the SOM model determined the highest number of candidates with the smallest length (Table 6, Table S11, S12, S13). Lastly, We found 122 gene and 14 lncRNA driver candidates common to five models in lung cancer, 103 gene and 12 lncRNA driver candidates in liver cancer, 1 gene and 0 lncRNA driver candidates in CLL and 305 gene and 18 lncRNA driver candidates in melanoma (Figure11, Figure S11, S12, S13 and Table S14, S15, S16). There was higher overlap among candidates from the 5 models as compared to random sampled ones (P=0 except lncRNA driver candidates for CLL, Table S17).

Table 6. General characteristics of PC gene and lncRNA driver candidates positively selected by each model in lung cancer

Tool	Adjusted P values < 0.05 Number of genes (Mb)		Average length (bp)	
	PCgene	LncRNA	PCgene	LncRNA
CADD	595(167)	180(17)	281945	96180
funSeq2	6779(679)	918(34)	100302	37671
GWAVA	2144(228)	1482(46)	103931	31476
SNP	7674(914)	976(52)	119191	54252
SOM	12797(1018)	10403(249)	79601	24001
Combined	11417(887)	5716(162)	77693	28443
Total Genes	20300(1266)	38263(456)	62412	11917

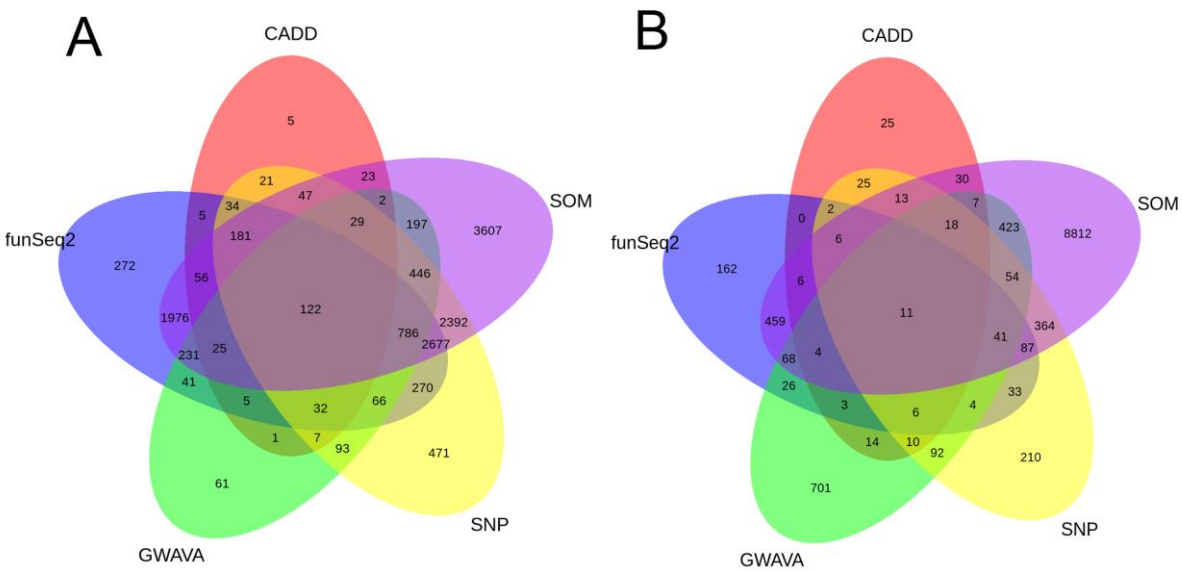

Figure 11. Comparison of driver candidates detected by five independent permutation models. A. Overlap of the driver gene candidates predicted by the 5 permutation models (CADD, FunSeq2, GWAVA, SNP and SOM); B. Overlap of lncRNA driver candidates predicted by the 5 permutation models (CADD, FunSeq2, GWAVA, SNP and SOM).

4.2.3 LncRNA driver candidates harboring enriched conserved elements

The evolutionary conservation of lncRNAs has been an ongoing subject of research, with several studies showing that lncRNAs are modestly conserved (Derrien, 2012, Necsulea, 2014, Guttman, 2010). We obtained evolutionarily conserved regions from the UCSC 46 mammalian genome alignment (Phastcons score >177) and mapped them onto lncRNA driver candidates. We performed a permutation test that randomly sampled regions with the same size as lncRNA drivers 1000 times from the whole lncRNAs set and computed the enrichment of conserved regions for each case. A P-value was produced by comparing observed enrichment of conserved elements with those of 1000 simulated samples.

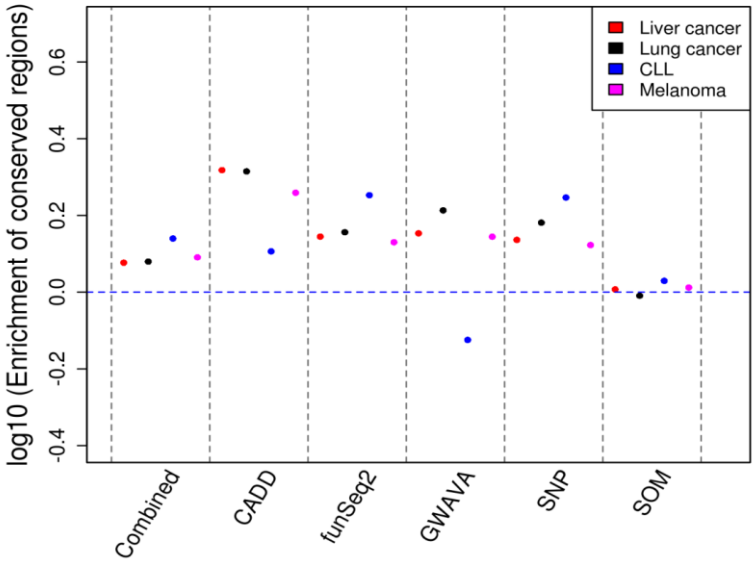

Figure 12. Enrichment for evolutionarily conserved regions within different lncRNA driver candidates in the four cancer types. For each feature, enrichment is computed as an odds ratio as explained in Methods. Values for each cancer are represented by a dot of distinct color. The blue dashed line denotes the baseline of enrichment of conserved regions in lncRNAs

Overall, the lncRNA predicted as positively selected by all models except SOM harbored higher enrichment for conserved regions than the random samples (P value <0.05 in all cases, Figure 12, Table S18). Owing to the large number of lncRNAs prioritized by the SOM model, these candidates showed similar level of enrichment for conserved regions as random samples (P value > 0.05 in three cases, a permutation test, Figure12, Table S18). For instance TTN-AS1 and HOXA-AS2, two lncRNAs which are positively selected by all models in lung cancer and show 41.05% and 42.54% of coverage of conserved regions respectively. In addition, these two lncRNAs are intensively overlapping with non-coding functional features, such as Dnase I hypersensitive clusters, H3K27ac, suggesting their function importance in lung cancer (Figure 13 - 14).

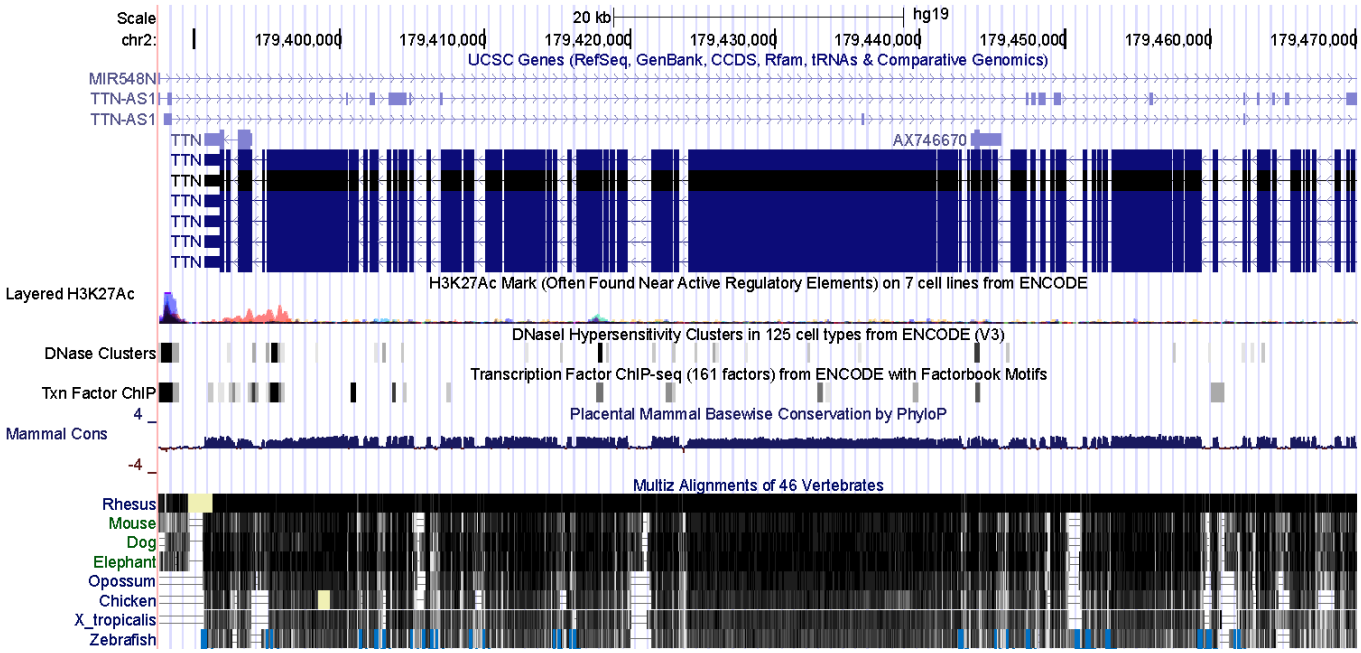

Figure 13. Graphical display of functional features in lncRNA TTN-AS1 from Genome browser

Mammal cons: conserved regions, Dnase Clusters: Dnase I hypersensitive clusters, Layered H3K27ac: H3K27ac

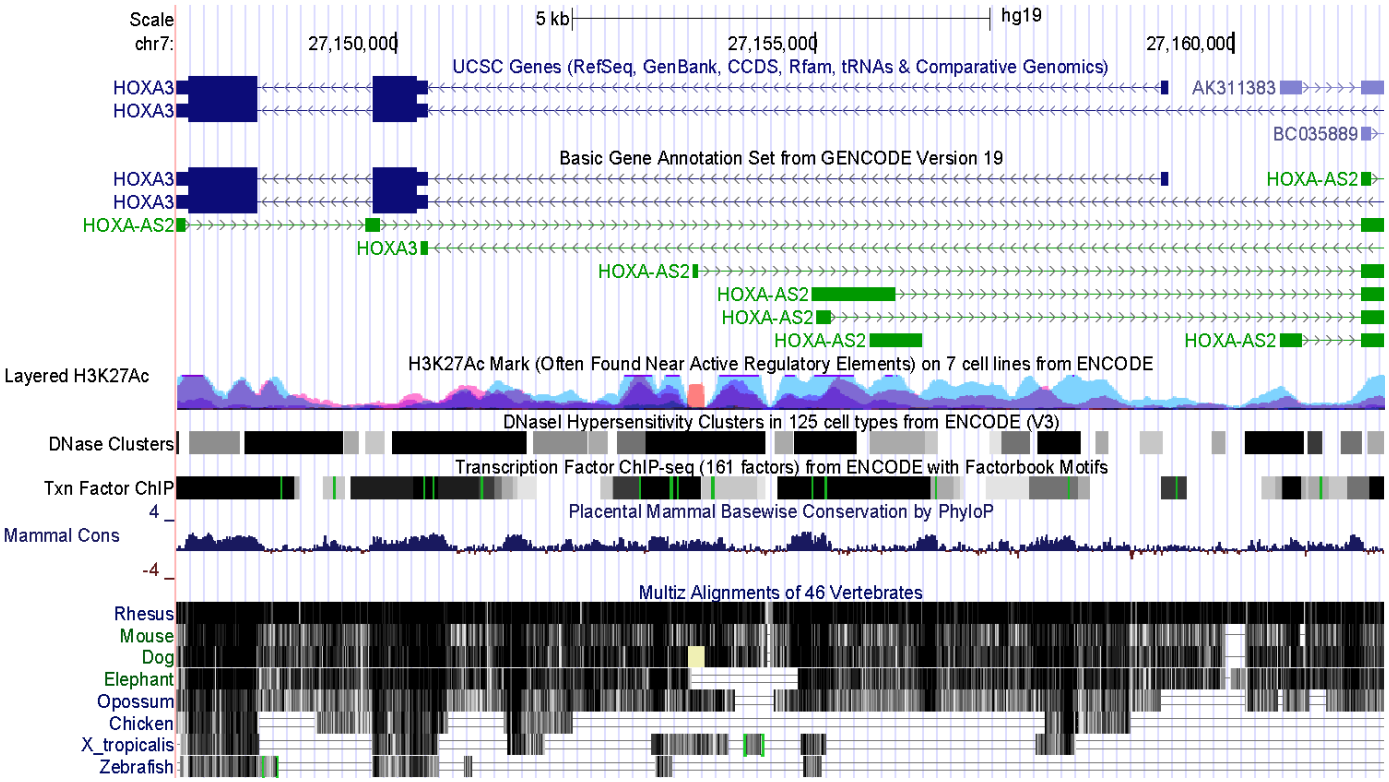

Figure 14. Graphical display of functional features in lncRNA HOXA-A2 from Genome browser (same legend as in Fig 13).

4.2.4 LncRNA driver candidates enriched for disease-associated variants

In order to further assess the functional importance of our lncRNA driver candidates, we analyzed their enrichment for HGMD and Clivariant disease-associated non-coding variants with the same permutation test as we did for the conservation analysis. Overall, we found 11/24 cases showing significantly increased enrichment for HGMD disease mutations compared to the random samples (P value <0.05, a permutation test, Figure15A, Table S19). Moreover, significant enrichment for Clivariant disease-associated variants was observed for 17/24 lncRNA driver candidates (P value <0.05, a permutation test, Figure15B, Table S19). These results suggest that, to a large extent, our lncRNA driver candidates are enriched for non-coding disease-causing variants and further support their functional importance in the non-coding cancer genome.

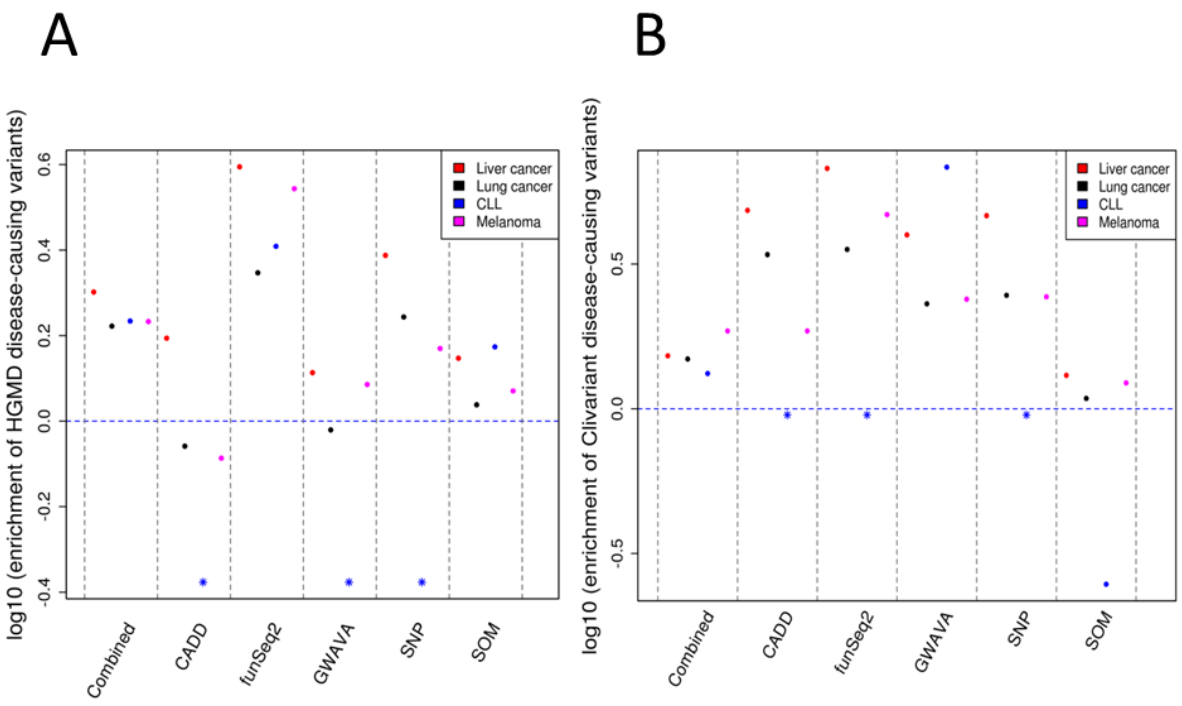

Figure 15. Enrichment for HGMD (A) and Clivariant (B) disease-causing variants within different lncRNA driver candidates in the four cancer types. For each feature, enrichment is computed as an odds ratio as explained in Methods. Values for each cancer are represented by a dot of distinct color. The blue dashed line denotes the baseline of enrichment of disease-causing variants in lncRNAs. The asterisks represent lncRNA driver candidates don't have HGMD and Clivariant disease-causing variants, their enrichment values are calculated as log10 (0.4202) and log10 (0.9524) respectively.

4.2.5 Expression analysis of lncRNAs in lung cancer

To further reduce the scope of screened cancer lncRNAs, we obtained RNA-seq data of normal lung and 85 cancer samples from Ju et al. (2012). 2208 lncRNAs were determined by DESeq2 Release (3.0) (Love et al., 2014) as differentially expressed between tumor and normal lung tissues with cutoffs of false discovery rate (FDR) ≤10e-4 and absolute fold change ≥2 (Figure 16, see methods). Among differentially expressed lncRNAs, 5 CADD, 45 funSeq2, 93 GWAVA, 54 SNP, 605 SOM and 335 combining drivers are differentially expressed between cancer and normal lung tissues. This list of lncRNAs will be potential driver candidates for experimental validation in lung cancer cells.

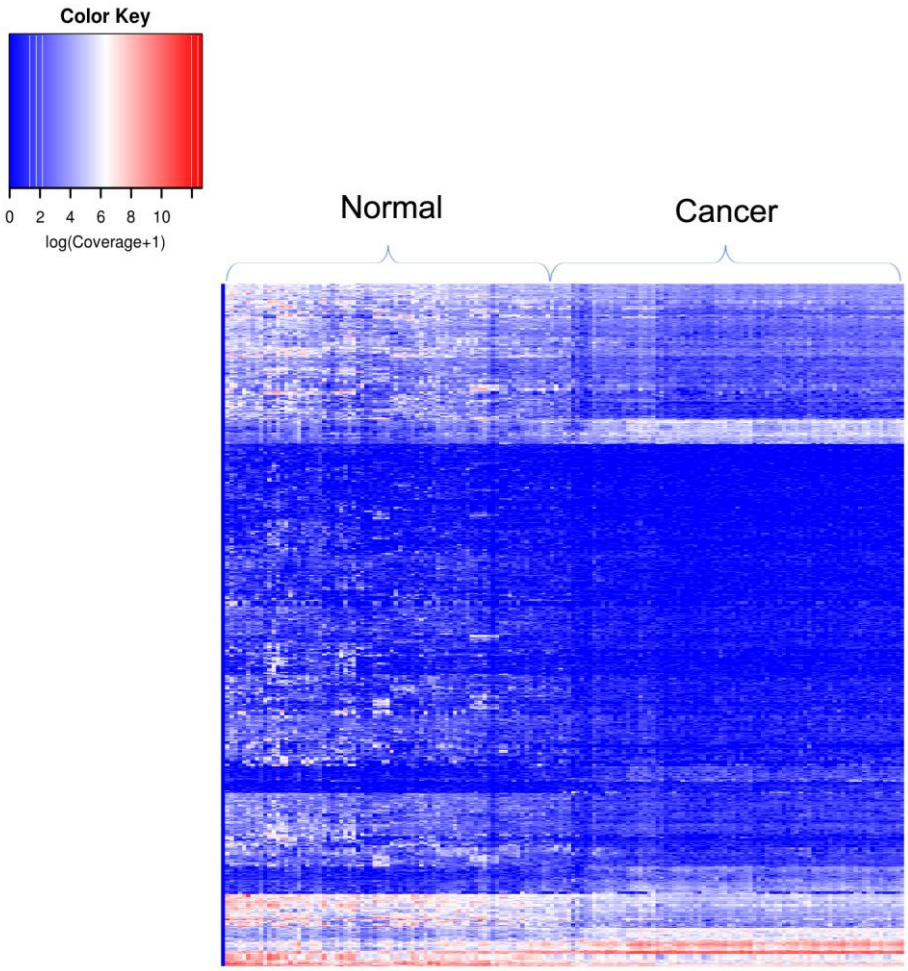

Figure 16. Heatmap showing normalized abundance of 2208 lncRNAs differentially expressed between lung cancer and normal lung tissues.

4.3 Discussion

The prioritization of cancer-associated lncRNAs is always a challenging and difficult task, as the mechanisms by which lncRNAs function are diverse and complex, ranging from gene transcription regulation, interaction with microRNAs or proteins to alternative splicing (Gutschner,2012). Over the past decade, researchers preferentially focused on lncRNAs which showed strong expression correlation with surrounding protein coding genes or interactions with proteins or miRNAs. A handful of computational tools have been developed to clarify lncRNA functions. However, little attention has been paid to the functional impact of non-coding mutations within lncRNAs and their importance to interpret cancer-associated lncRNAs. In this study, we tried to resolve this problem based on a permutation-based model which screens potential cancer-associated lncRNAs displaying a shift towards accumulation of non-coding variants with high functional impact. We applied the model to both cancer genes and lncRNAs using their non-coding somatic mutations in 4 cancer types, the results obtained showed that both cancer genes and lncRNAs have significantly lower adjusted P values than generic protein coding genes and lncRNAs in all cases, strongly supporting the validity of our model. As demonstrated in the Gonzalez-Perez et al 's study, the coding regions of cancer drivers preferentially accumulate mutations with high functional impact, an important concept that we carry out further in this study by showing this functional bias is absolutely applicable to non-coding regions such as UTRs or introns. Most importantly, despite their lack of coding potential, cancer lncRNAs exhibit the same trend.

In addition, we carried out a permutation model on the whole lncRNA dataset and obtained hundreds of cancer-related lncRNA candidates. Further characterization of these lncRNAs showed they are a subset of lncRNAs enriched for evolutionary conserved regions and disease-associated variants, highlighting their functional importance. We listed a handful of lncRNA candidates, such as TTN-AS1, HOXB-AS3 and HOXA-AS2. Not only do these lncRNAs contain high coverage of evolutionarily conserved regions, but also they are intensively overlapping with non-coding functional features, such as Dnase I hypersensitive clusters and open histone marks. The lncRNA HOXA cluster antisense RNA 2 (HOXA-AS2), located between the HOXA3 and HOXA4 genes, has been functionally characterized in

leukemia (Zhao et al., 2013) and gastric cancer (Xie et al., 2015). The knockdown of its expression reduced cell viability and induced cell apoptosis in NB4 promyelocytic leukemia cells possibly through TNF-related apoptosis-inducing ligand (TRAIL) pathway (Zhao et al., 2013). Moreover, HOXA-AS2 is aberrantly expressed and plays an oncogene role in gastric cancer, knockdown of HOXA-AS2 markedly suppressed gastric cells growth by initiating G1 arrest and enhancing apoptosis in part through inhibiting P21, PLK3, and DDIT3 expression (Xie et al., 2015). However, further experimental validation is still needed for other cancer lncRNA candidates to characterize their functional roles in cancer.

There still a lack of efficient bioinformatics tools to prioritize cancer-related lncRNAs on a whole genome scale. A contribution of this work is that it might greatly reduce the scope of screening cancer lncRNAs for oncology researchers, simply based on the mutation pattern and function information of non-coding mutations within lncRNAs. However, many concerns still exist, for example, the SOM scores are computed on a 1-Kb scale, the other 4 scoring systems have a nucleotide-level scoring precision, which leads to a large number of positively selected lncRNAs by SOM model and greatly increases false positive rate, therefore an improvement is still needed with respect to increasing the prediction accuracy of the SOM model and reducing the number of false positively selected lncRNAs. Alternatively, we could find a way to combine the SNP and SOM scores to form an uniform score and then use it in the permutation test. These will be our objectives in the future.

4.4 Methods and materials

4.4.1 Cancer mutation, disease-causing variants, lncRNAs and cancer gene and lncRNA data

Somatic variants were collected from whole genome sequencing of paired cancer and normal tissues, obtained from two studies: 2,011,261 variants from 25 melanoma patients (Berger et al., 2012), 1,845,976 from 24 lung adenocarcinoma patients, 881,136 from 88 liver cancer patients and 59,993 from 28 chronic lymphocytic leukemia (CLL) patients (Lawrence et al., 2013). Variants described as "substitution" or "indel" were both collected and are referred to collectively as mutations in the text.

Curated disease-related variants were obtained from the Clivariant (Version 2014/03/03, 55,689 variants) (Landrum et al., 2014) and HGMD (Version 2014/04/14, 166,768 variants) databases (Stenson et al., 2009). After exclusion of coding positions we used 13,108 HGMD and 6045 Clivariant mutations.

LncRNA annotation mainly comes from three different sources, Gencode v7 (Harrow J, 2012), Human Body Map lincRNAs (large intergenic non coding RNAs) and TUCPs (transcripts of uncertain coding potential) generated from 4 billion RNA-Seq reads across 24 tissues and cell types (Pj et al., 2012) as well as Refseq annotation (Pruitt et al., 2007). In total, there are 38263 lncRNA annotations (456.01 Mb) collected from these three different databases. Lists of cancer genes were obtained as follows: cancer-related lncRNAs are 61 mammalian long non-coding transcripts identified from our literature search as experimentally associated with different cancer types (Table S20); protein-coding cancer genes are from the Cancer Gene census, available from COSMIC release V71 (http://cancer.sanger.ac.uk/cancergenome/ projects/census/) (Forbes et al., 2011a).

4.4.2 Scoring non-coding variants

In total, non-coding variants were scored using CADD (http://cadd.gs.washington.edu/), FunSeq2 (http://funseq2.gersteinlab.org/), GWAVA (https://www.sanger.ac.uk/sanger/StatGen_ Gwava) , SNP model and SOM models respectively for each cancer type, all the parameters were set to default. Of note, we used the "region" classifier of GWAVA which is trained using regulatory variants of HGMD and a random selection of SNVs from across the genome to measure function effect of non-coding variants.

4.4.3 The permutation-based model

The permutation-based model relies on the hypothesis that cancer-related lncRNAs display a bias toward accumulation of non-coding variants with high function impact. Take lncRNA A and lncRNA B as examples (Figure S14A): lncRNA A is more enriched with non-coding variants with high function impact as compared to lncRNA B, therefore, lncRNA A is more likely to be non_coding driver in cancer. The permutation-based model consists of two main steps. First, all non-coding variants are scored with CADD, FunSeq2, GWAVA, SNP model

and SOM model of lung cancer respectively, then the average scores are computed for each lncRNA based on the observed variants in that specific lncRNA; the second step is a permutation test to examine which lncRNAs exhibit a function impact bias. As for each lncRNA and scoring system, it randomly takes the same number of observed variants with replacement from all the non-coding variants found in all sequenced samples and computes the corresponding average score, this random sampling is repeated 1,000,000 times, generating a null distribution of average scores for each lncRNA and scoring system (Figure S14B). Empirical Pvalues represent the fraction of sampling average scores greater than the observed ones, however, as for the SOM score, P values refer to the fraction of sampling average scores less than the observed ones. The P values from five independent permutation tests are adjusted using False Discovery Rate (FDR) (Yekutieli and Benjamini, 1999). In addition, we also run the permutation-based model on PCgenes using their non-coding somatic mutations. Driver candidates are defined as PCgenes and lncRNAs whose adjusted P values are less than 0.05. Finally, we use a z transform (Whitlock, 2005) to combine five different P values of each PCgene and lncRNA to form an uniform P value.

4.4.4 RNA-seq data processing and expression analyses of lncRNAs

161 RNA-seq data including 76 normal lung samples and 85 cancer samples were obtained from Ju et al's study (Ju et al., 2012). Reads were mapped to the hg19 genome using Star aligner (Dobin et al., 2013). Read counts were computed with bedtools v2.22.1 for each lncRNA (Quinlan and Hall, 2010); DESeq2 Release (3.0) (Love et al., 2014) was used to identify differentially expressed transcripts between tumor and normal pairs with cutoffs of false discovery rate (FDR)(Yekutieli and Benjamini, 1999) ≤10e-4 and absolute fold change ≥2.

4.4.5 Enrichment analysis

Enrichment for conserved regions or HGMD and Clivariant disease-associated variants within different driver candidate classes (Fig 12 and 15) was measured as the odds ratio:

$$enrichment = \frac{\left(\frac{Hf}{Sf}\right)}{\left(\frac{Hg}{Sg}\right)}$$

Where Hf = size of conserved regions or the number of HGMD and Clivariant disease-

associated variants within driver candidate, *Sf* = total size of driver candidate, *Hg* = size of conserved regions or the number of HGMD and Clivariant disease-associated variants in whole lncRNAs, *Sg* = total size of lncRNAs. The significance of enrichment or depletion was evaluated using a permutation test as follows: a set of positions of same size as the driver candidate (ie. 17.31 Mb) was randomly sampled from the whole lncRNAs set 1000 times, and in each random sample, enrichments were calculated for each driver candidate class. Enrichment for HGMD and Clivariant disease-associated variants was evaluated similarly.

4.4.6 Statistical analyses

Data were presented as mean, differences between different groups were drawn with the Fisher exact test and Wilcoxon rank sum test in R, $P < 0.05$ was regarded statistically significant and the null hypothesis was rejected.

Chapter 5 - Conclusion and perspectives

5.1 General conclusion

Functional annotation of cancer mutations have been a consistent focus of cancer genomics studies. In the past, researchers preferentially focused on mutations in the coding fraction of human genome. Ample bioinformatics tools have been developed to distinguish cancer-driver mutations from neutral ones, such as SIFT, polyphen2 and MutationAssessor. As descried in detail in the introduction, these tools can be classified as three main groups, empirical, machine learning and hybrid approaches. The rationales of these programs lie in a variety of properties ranging from evolutionary conservation, physicochemical constraints, protein structures and curation of disease-associated mutations. Based on function information of coding mutations, the downstream work is searching for cancer driver genes that are critical to cancer formation and progression. The most common approach (ie MutSigCV and MuSiC) detects recurrently mutated genes as cancer-driving. However, as cancer drivers can also occur at a low frequency, new programs independent of cancer mutation frequency have been developed (ie Oncodrive-fm, OncodriveCLUST and InVEx).

In recent years, as an increasing number of variants have been identified as disease-associated in the non-coding genome, interpreting non-coding cancer mutations has become an urgent task in cancer genomics studies. The completion of large projects, such as ENCODE, has made functional interpretation of cancer variants achievable. Multiple programs have been built based on this functional information. As described in the introduction part, these tools can be divided into empirical approach such as RegulomeDB, funSeq2 and machine learning model such as CAAD and GWAVA. In Chapter 2 of this study, in order to functionally interpret non-coding mutations in cancer and eventually identify new cancer drivers, we took into account the dual selection forces acting on the tumor genome: (1) population and evolutionary constraints acting at germline level and (2) constraints resulting from the accelerated mutation background of the cancer tissue. To achieve this, we have developed two independent models, refered to as SNP and SOM models. Given a combination of features, the SNP model was constructed to predict expected fraction of rare SNPs using random forest model, the second SOM model was built to compute the expected mutation density for each 1-Mb window with an array of feature types ranging from replication time, expression level, histone modifications to regulatory elements. We applied our two models to score these

disease-associated variants and a set of random control SNPs. Our results showed that the SNP and SOM models are capable of distinguishing Clinvariant and HGMD disease-causing mutations from neutral ones. In addition, we intersected high SNP scoring and low SOM scoring regions and obtained 56 Mb functionally important regions (referred to as hypomutated regions). This small portion of the human genome shows highest enrichment of disease-causing variants among intergenic, low SOM scoring, high SNP scoring and hypomutated regions, further supporting low somatic mutation areas and high ratio of rare SNPs regions are functionally relevant and can be used as a screen for prioritizing cancer-related non-coding mutations. This study demonstrated that purifying selection as measured by fraction of rare SNPs and mutation density constraints are informative for the evaluation of functional impact of cancer mutations in the non-coding genome. Moreover, combination of the SNP and SOM models would fasciliate the prediction of disease mutations in the non-coding genome.

Another important part in my thesis (Chapter 4) was the application of the scoring tools CADD, funSeq2, GWAVA and our SNP and SOM scoring systems to prioritize cancer-associated lncRNAs with a permutation-based algorithm. We hypothesized that accumulation of non-coding mutations with high function impact indicates a positive selection in cancer genome and cancer-related lncRNAs show a bias toward enrichment of high functional non-coding variants. We implemented the permutation model on 61 cancer-related lncRNAs and 452 cancer genes using cancer mutation data of liver cancer, lung cancer, CLL and melanoma. We observed that both cancer lncRNAs and genes had lower average adjusted P values than total lncRNAs and genes. These results suggest that cancer-related lncRNAs and genes are enriched for non-coding variants with high functional impact. Applying the permutation test to lncRNAs with five different scoring systems enabled us to prioritize hundreds to thousands cancer-related lncRNA candidates. We would recommend to combine the adjusted P value and ranking of the P value to prioritize potential cancer-related lncRNA candidates. Furthermore, if we focus on those lncRNA candidates which are positively selected by all five scoring systems, the number of cancer-related lncRNAs candidates could be reduced to 11 in lung cancer, 11 in liver cancer, 0 in CLL and 18 in melanoma. These lncRNA candidates can be used for experimental validation. For example, we could study their function role in cancer

cells via over-expression or silencing.

Taken together, we have successfully developed two models, SNP and SOM, to measure the functional impact of non-coding variants in the cancer genome. Injecting these scoring systems to a permutation-based model enables us to prioritize cancer-associated lncRNAs on a genome scale. The completion of our project paves the way for further characterization of unknown cancer mutations and lncRNAs in the non-coding cancer genome.

5.2 Perspectives

5.2.1 Refinement of the SOM and SNP models

Due to the sparse number of cancer mutations, the SOM model was built based on a 1-Mb window and the SOM scores were computed and averaged on a 1-Kb scale. As more and more whole genome sequencing studies are ongoing, there will be an explosive increase in the number of publically available cancer mutations, which should enable us to construct the SOM model with 1-Kb window and should remarkably improves the prediction accuracy. In addition, an increased prediction accuracy of the SOM score will greatly reduce the number of cancer-related lncRNA candidates positively selected by the SOM model and diminish the false positive rate. Lastly, as an accumulating number of new functional features are produced, adding these features to the SNP and SOM models and retraining the two models will further refine their prediction capability.

5.2.2 Integrating SNP and SOM scores to form a combined score

As shown above, there exists a remarkable difference between the SNP and SOM scoring systems with respect to score range and the importance of predicting disease-causing variants. As we did not find a satisfying way to integrate the two scores, an important work for us will be to come up with a way to combine scores and apply the combined score to the permutation-based model, which should reduce the number of cancer-related lncRNA candidates prioritized by the SOM model and its negative impact on the combined P value in the future.

5.2.3 Functional analysis of cancer lncRNA candidates

We have identified a list of cancer-specific lncRNAs as prioritized by our permutation-based model. There is plenty of future work on this basis, such as analyzing where these lncRNAs are expressed, what expression levels they have in cancers and what relations they have with surrounding genes. Most importantly, in order to clarify the functional potential of the positively selected lncRNAs, experimental validation is needed. Ectopic expression using lentiviral vector and siRNA-mediated knockdown should be conducted in cancer cells. Their effect on cellular proliferation, apoptosis and metastasis should be examined through MTT, flow cytometry, transwell and wound-healing assays, respectively. This work will be performed with our collaborators in Institut Curie.

5.2.4 Setting up an user-friendly website

The objective of our project was to develop a scoring system for measuring the function impact of non-coding variants and provide a program to screen the ncRNA transcriptome for potential cancer-associated lncRNAs based on somatic mutations. These goals have been in part achieved, but an important future work will be to construct a user-friendly interface, to enable submission of somatic mutation data and obtain their SNP and SOM scores. Moreover, users may also upload mutation data generated by whole genome sequencing of a cohort of cancer samples and obtain a list of potential cancer-related lncRNAs for further experimental characterization.

Chapter 6 - Appendix

6.1 Supplementary Figures

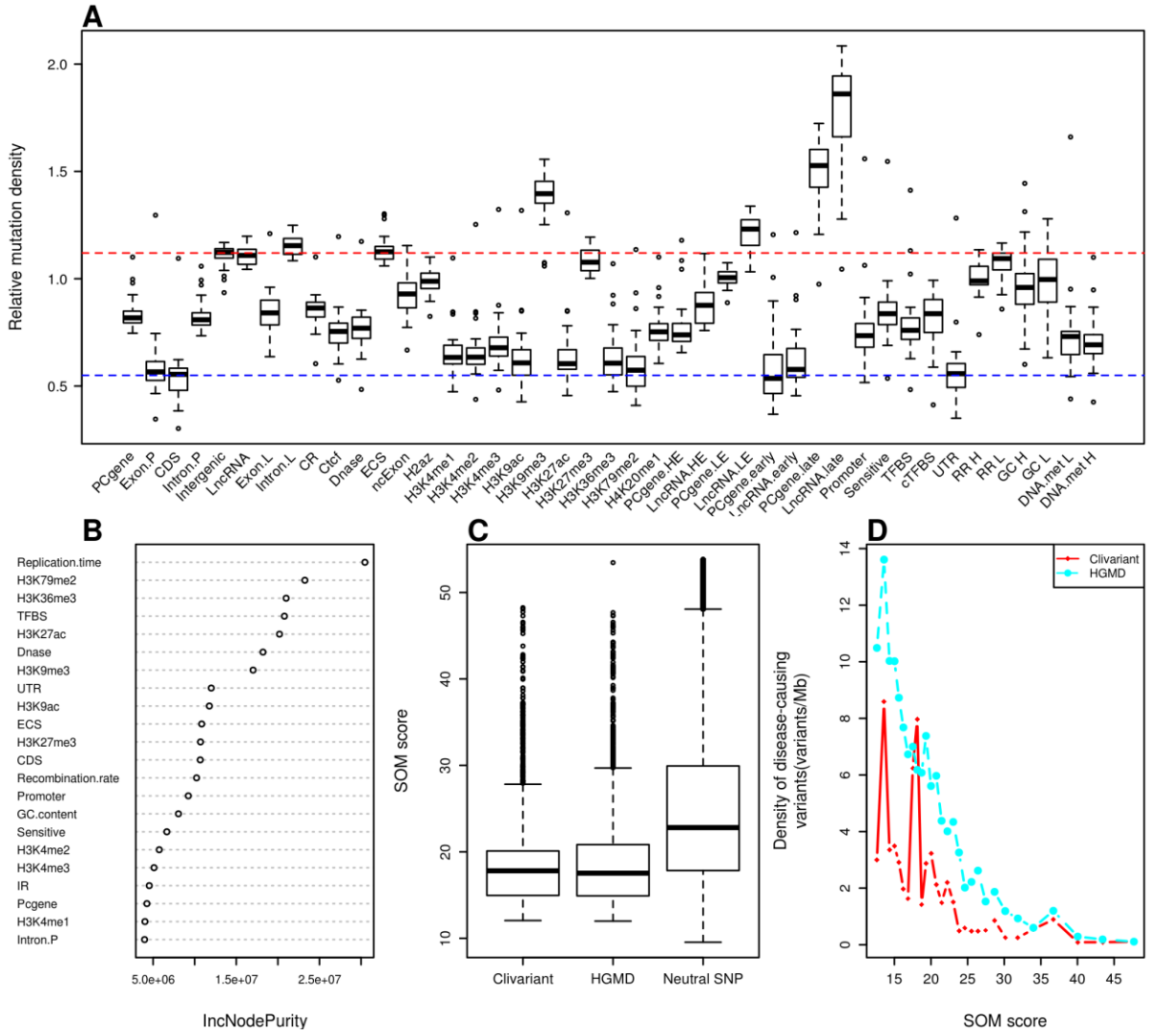

Figure S1. Construction of the Somatic Mutation (SOM) model for lung cancer. **A**. Relative density of somatic mutations from whole genome sequences of 24 lung cancer, associated to different genome features (see Methods for feature details). Mutation density is normalized so that the whole genome average has a mutation density of 1. PCgene: protein coding gene; CDS: coding sequence; Exon.P , Intron.P, Exon.L,Intron.L are exon and intron of protein coding gene and lncRNA respectively; CR: conserved region; DNase: DNase I hypersensitive site; ECS: evolutionarily conserved structure; ncExon: non-coding exon; PCgene.HE, LncRNA.HE, PCgene.LE and LncRNA.LE are high expressed and low expressed protein coding gene and lncRNA; PCgene.early, LncRNA.early, PCgene.late and LncRNA.late are early and late replicated protein coding gene and lncRNA; cTFBS: conserved

transcription factor binding site; RR H,RR L,GC H,GC L,DNA.met H and DNA.met L are 1-Kb windows with high recombination rate (> 4.0), low recombination rate (< 0.5), high GC content (GC % > 50%), low GC content (GC%<30%), high DNA methylation (average value > 0.7245) and low DNA methylation (average value < 0.4062) respectively; Red and blue dotted lines: base lines from CDS and intergenic regions; **B:** Feature importance as measured by IncNodePurity. We only show here features that passed feature selection. **C**. Distribution of SOM scores for neutral SNPs and for clinical variants from two disease-causing variants databases Clivariant and HGMD. Neutral SNPs here are the SNPs with allele frequency higher than 0.01 from the 1000 Genome project, SOM scores were predicted by the random forest model and divided by the number of patients. **D**. Correlation of SOM score with densities of disease-causing variants. The purple dotted line shows cutoff used for defining low SOM score in lung cancer.

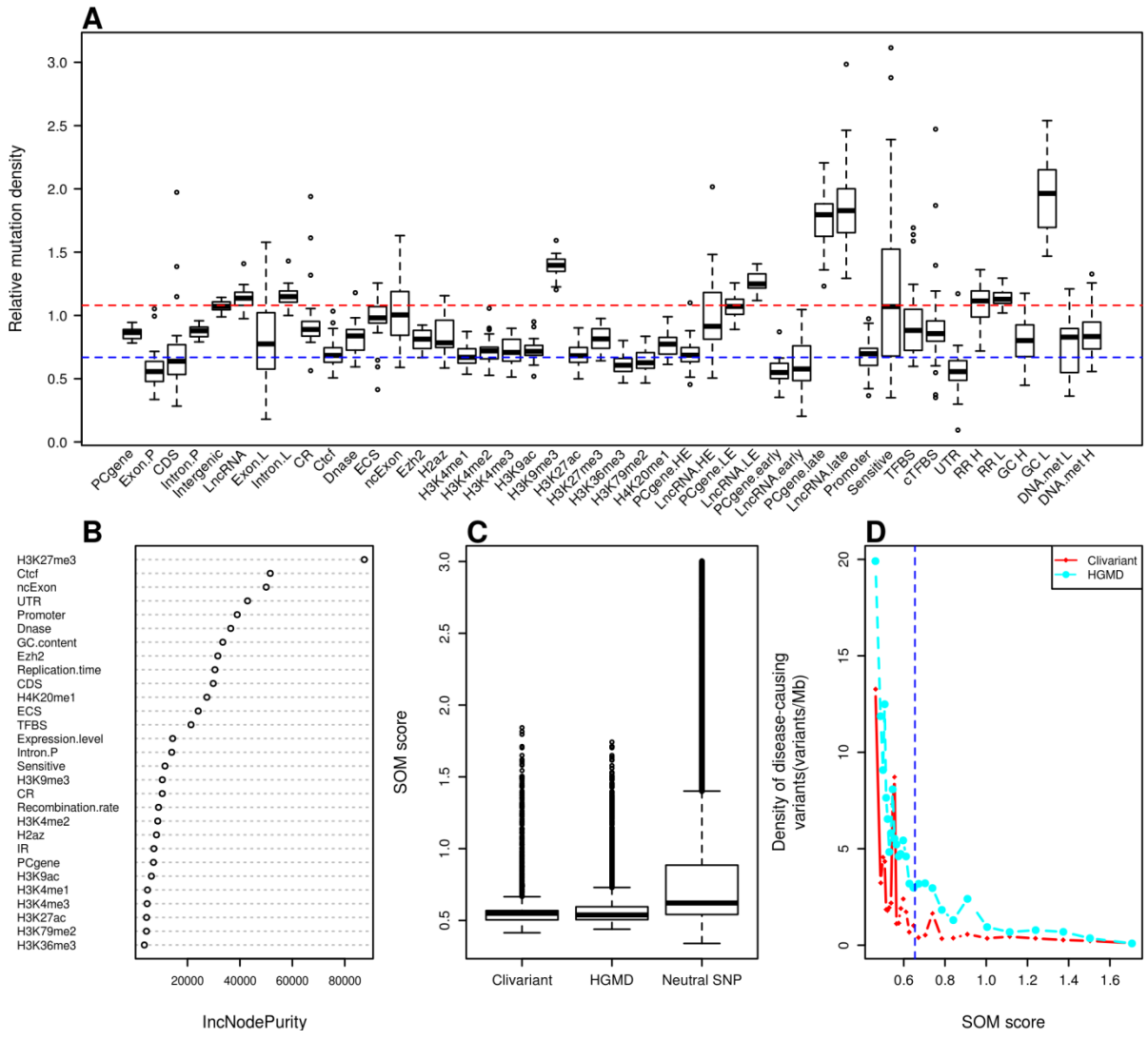

Figure S2. Construction of the Somatic Mutation (SOM) model for CLL. See Fig S1 for legend.

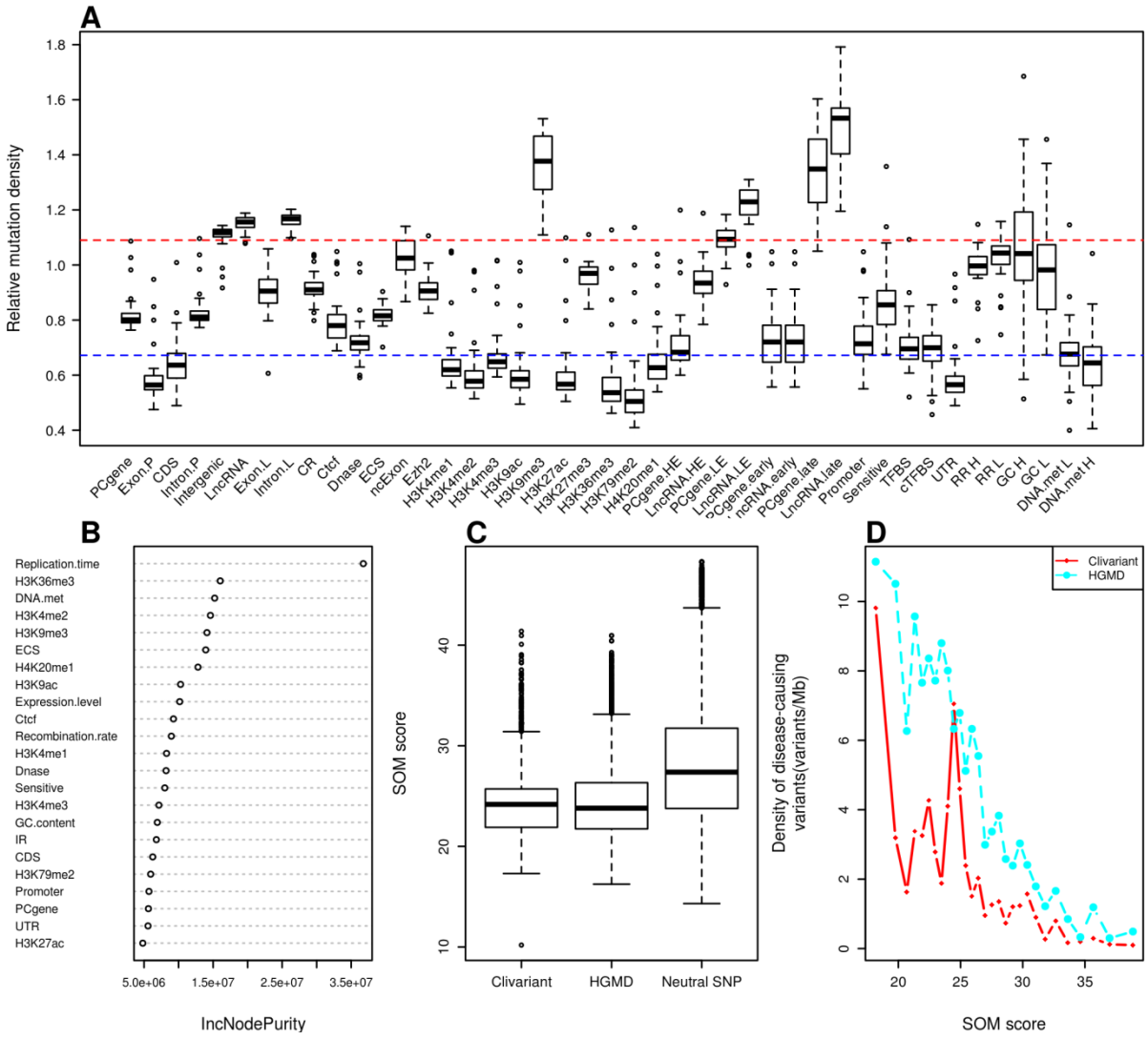

Figure S3. Construction of the Somatic Mutation (SOM) model for melanoma. See Fig S1 for legend.

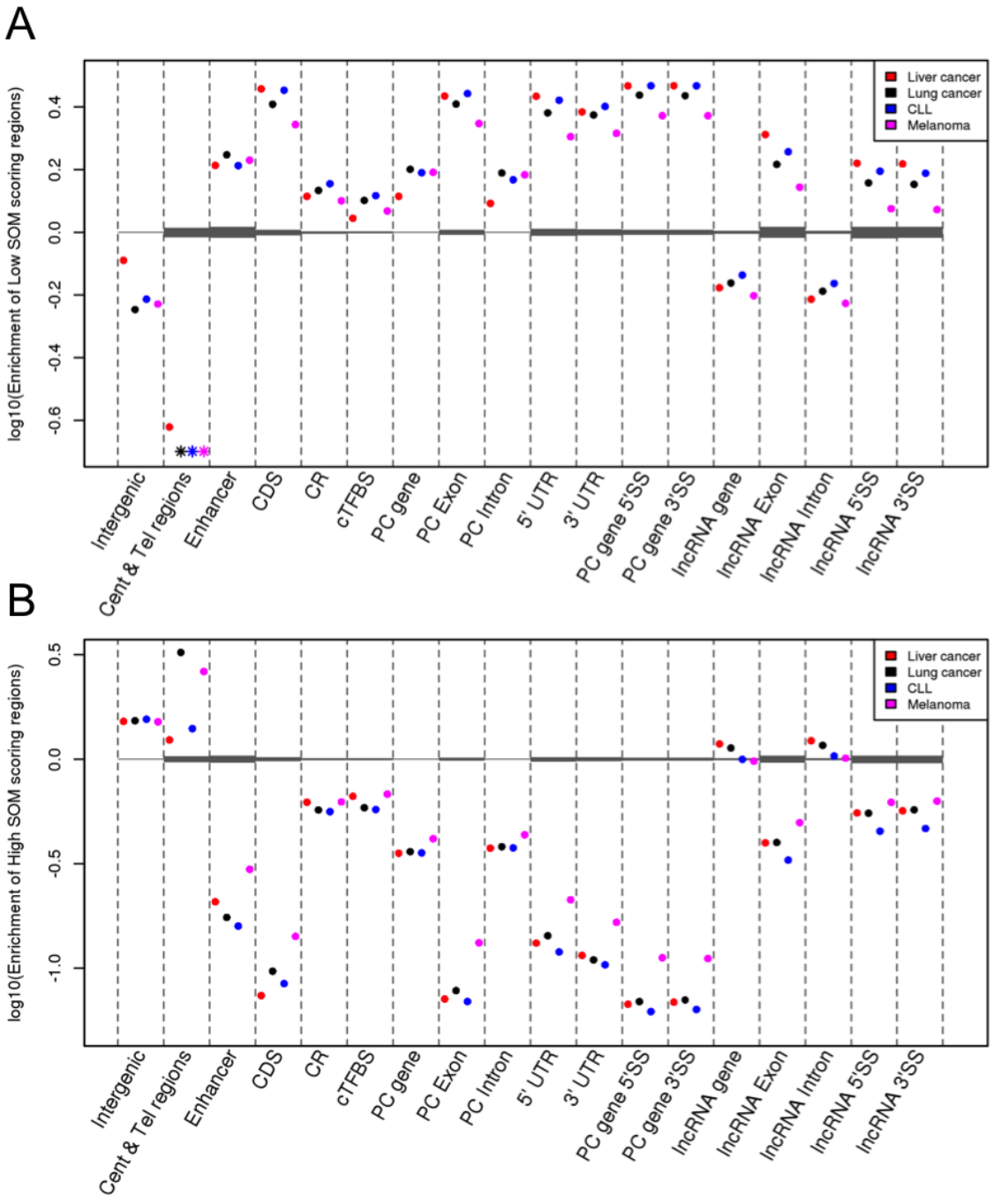

Figure S4. Enrichment for low SOM score (A) or high SOM score (B) positions within genome features in the four cancer types. Low (high) SOM score regions are defined as the 300M positions of the genome with lowest (highest) SOM score. For each feature, enrichment is computed as an odds ratio as explained in Methods. Shaded grey areas show enrichment ranges obtained from 1000 random permutations of the 300M positions (see Methods). Values for each cancer are represented by a dot of distinct color. The asterisks represent genome feature (Cent & Tel regions) doesn't overlap with low SOM regions, their enrichment values are calculated as log10 (0.2).

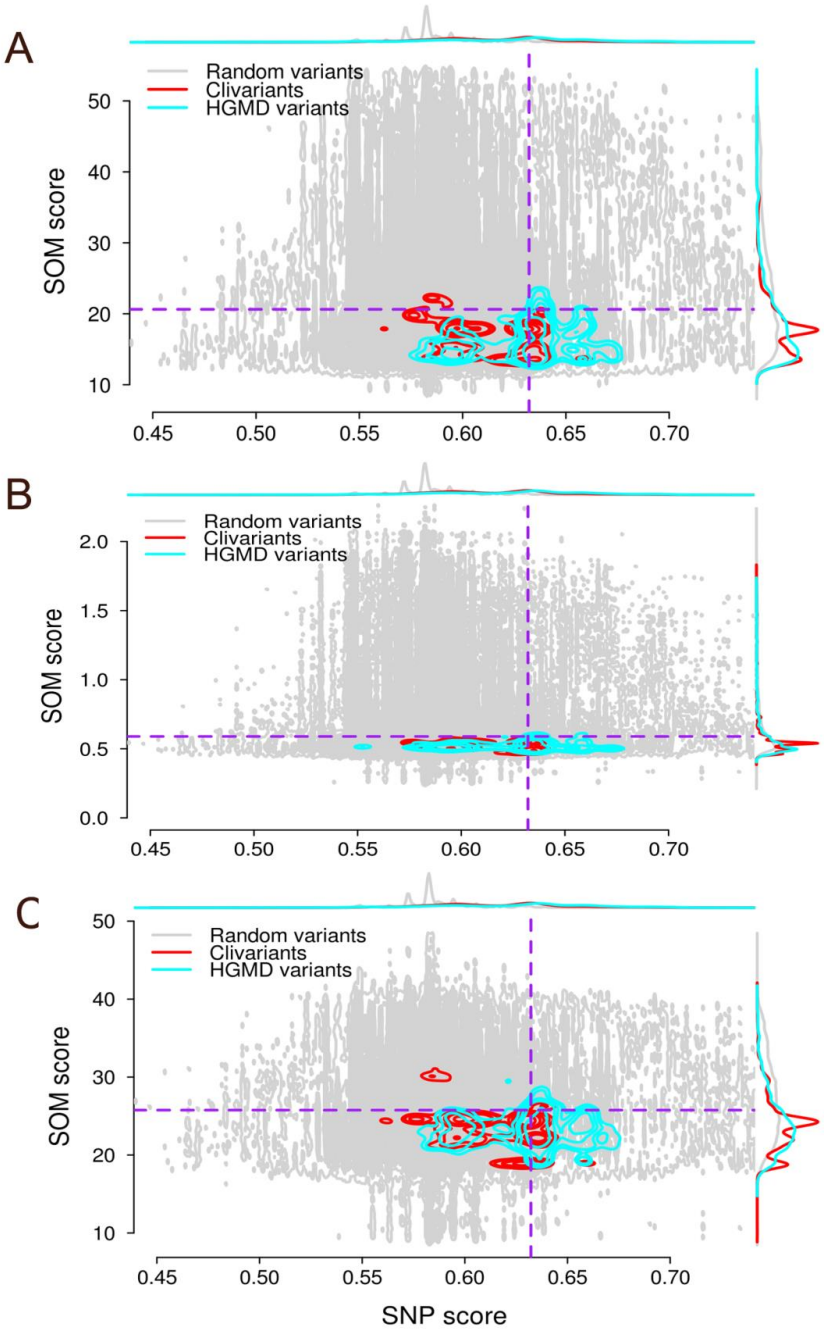

Figure S5. Relationship between SNP and SOM scores in lung cancer (A), CLL (B) and melanoma (C). Grey dots: 1 million random genome positions; cyan contour: HGMD disease-causing variant positions; red contour: Clivariant positions. The top and right curves show marginal distributions of SNP scores (top) and SOM scores (right) for random genome positions, HGMD and Clivariant disease-causing variants. SNP score cutoff=0.63 (100Mb above cutoff), SOM score cutoffs = 20.63, 0.59 and 25.76 variants/Mb, defining areas below cutoff of 1186.45 Mb, 1236.51Mb and 1170.98Mb in lung cancer, CLL and melanoma, respectively. Hypomutated regions (bottom, right area) defined by

both cutoffs correspond to ~56Mb in each cancer type.

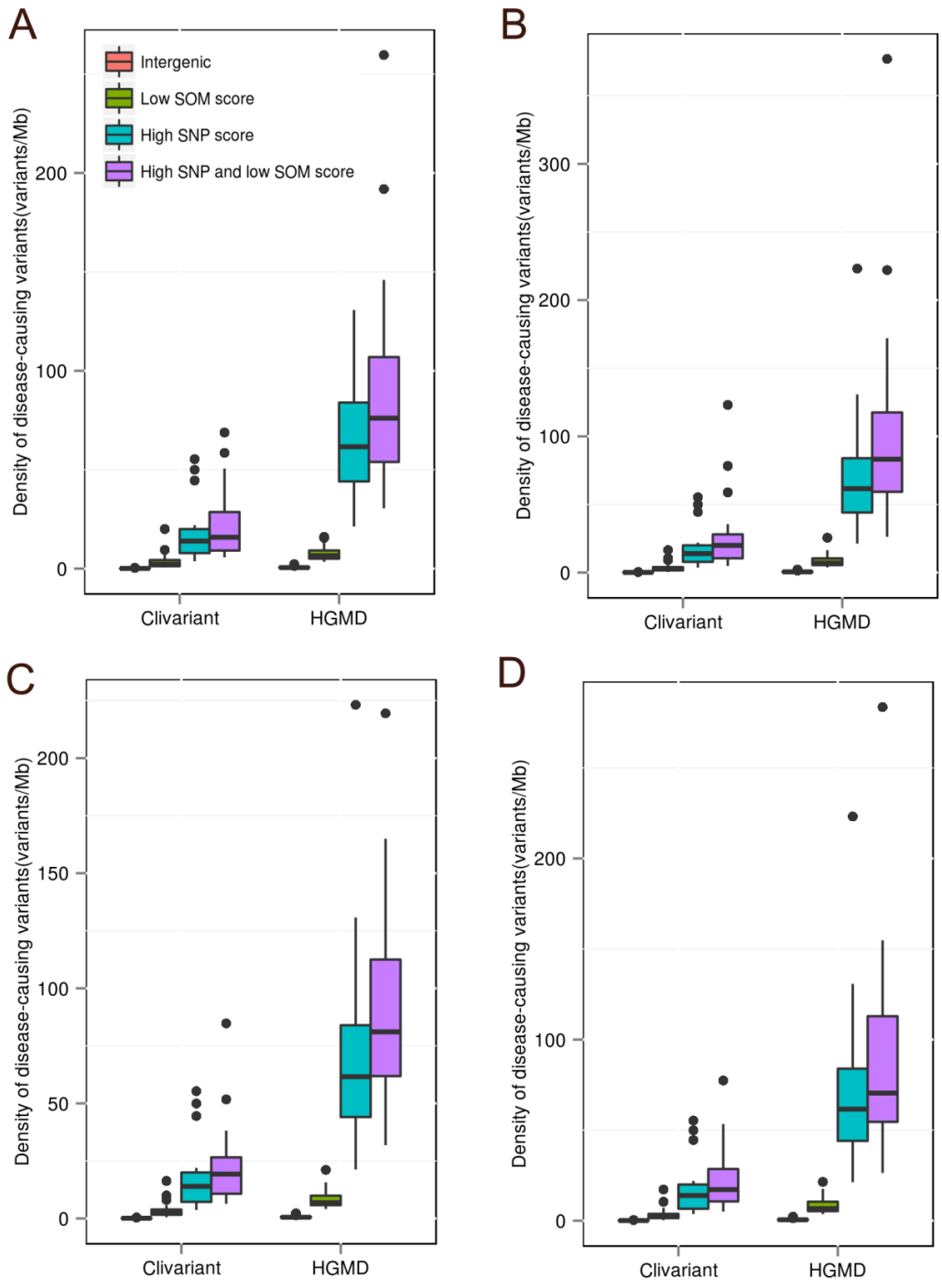

Figure S6. Effect of combining high SNP scores and low SOM scores in 4 cancer types (A: liver cancer, B: lung cancer, C: CLL, D: melanoma). For each chromosome, the size of intergenic, high SNP, low SOM and high SNP + low SOM regions, was calculated and numbers of disease-associated variants either from HGMD or Clivariant were counted. The boxplot shows densities of disease-

associated variants in each type of region, chromosome by chromosome. Cutoffs for defining high SNP and low SOM are the same as in Fig 3.

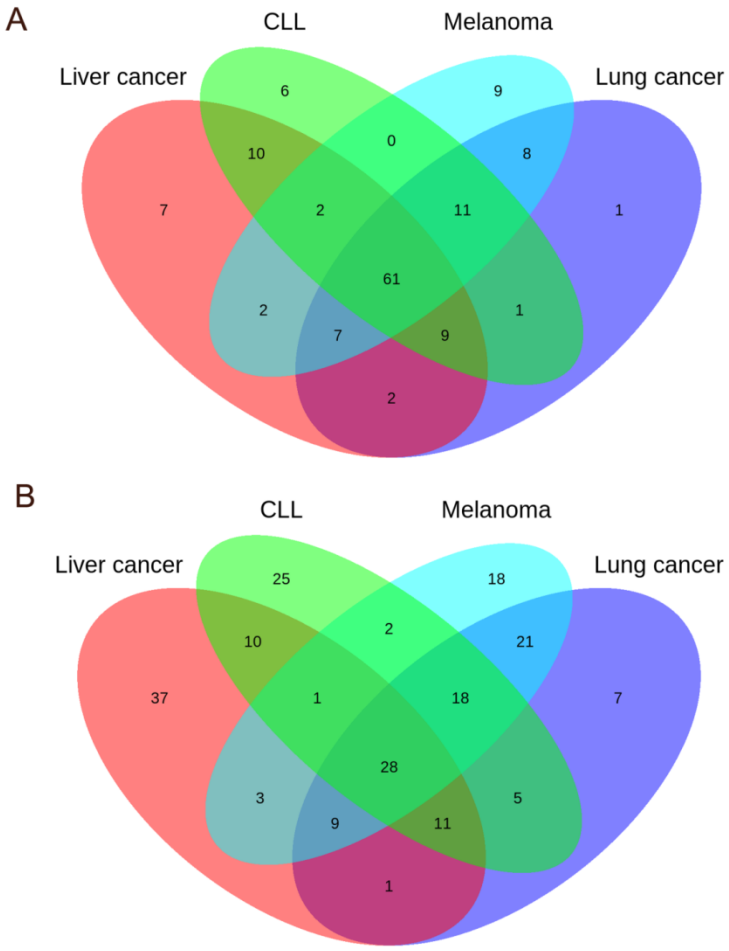

Figure S7. Venn diagrams showing the distribution of genes covered by hypomutated (A) or hypermutated (B) positions, across the 4 cancer types. In each cancer type the 100 genes with the highest coverage by hyper/hypomutated regions is shown.

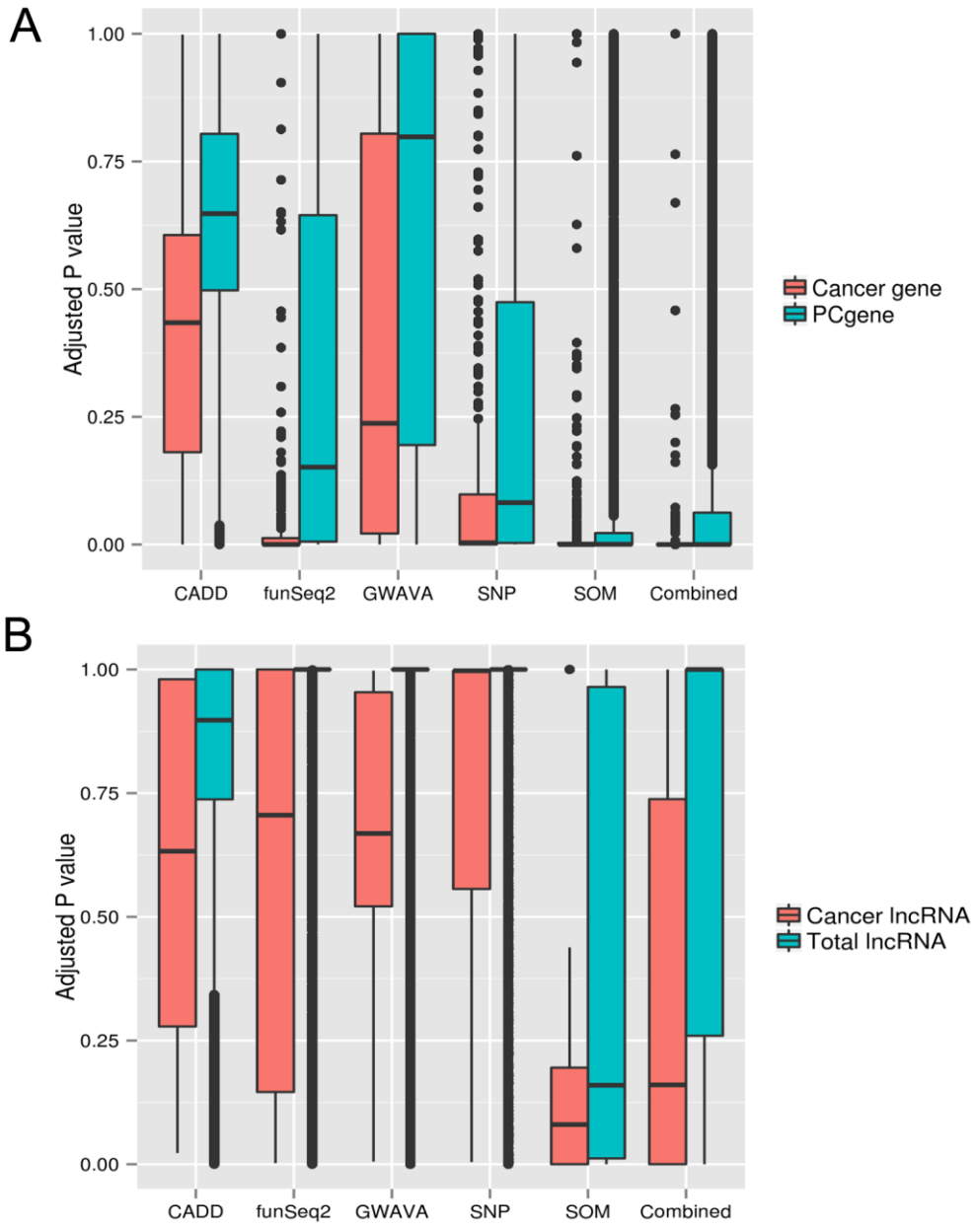

Figure S8. Distribution of adjusted P values for different gene classes in liver cancer. A. The comparison of adjusted P values computed by all permutation models between cancer-related genes and all genes; B. The comparison of adjusted P values computed by all permutation models between cancer-related lncRNAs and all lncRNAs.

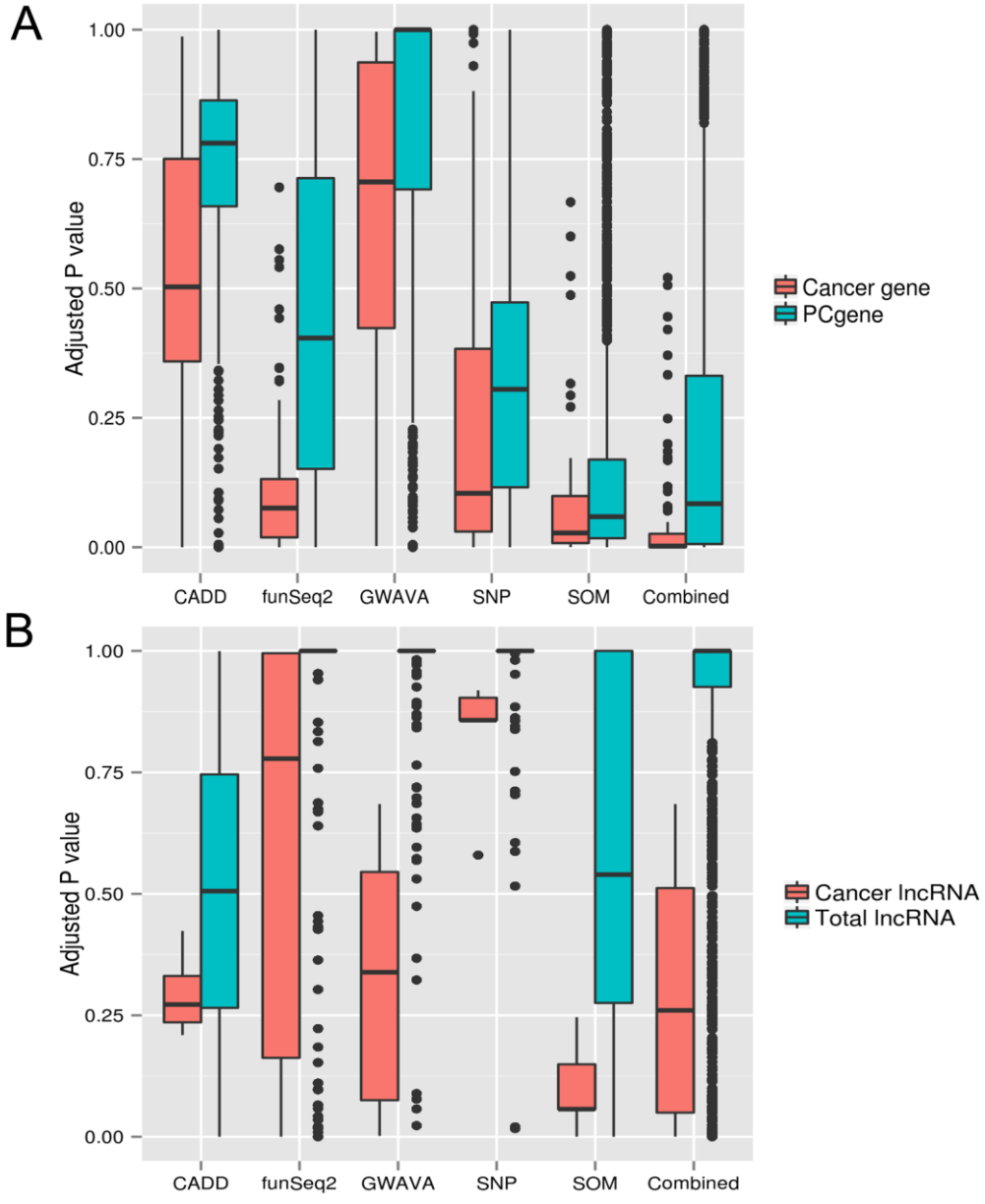

Figure S9. Distribution of adjusted P values for different gene classes in CLL. See S8 for legend.

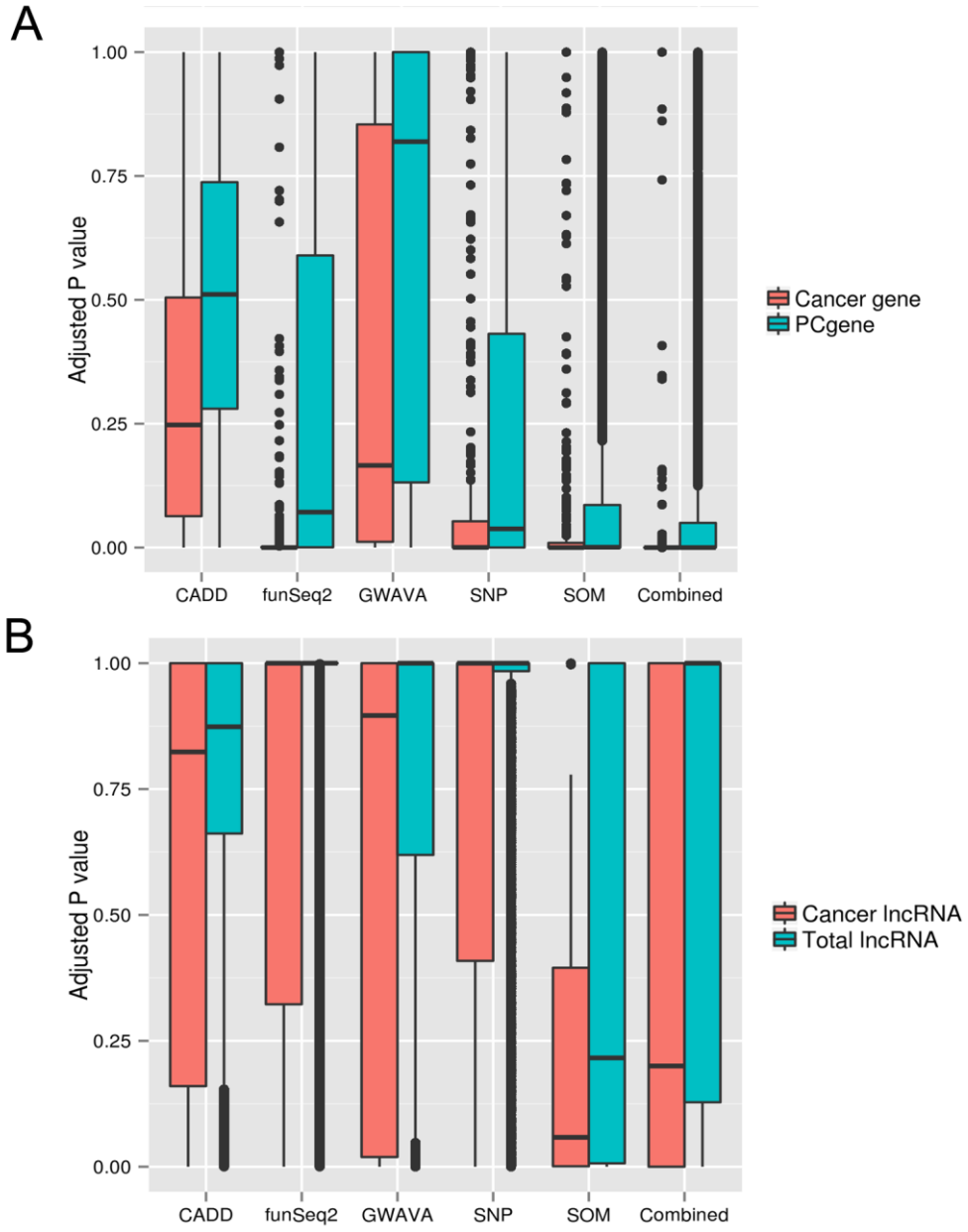

Figure S10. Distribution of adjusted P values for different gene classes in melanoma. See S8 for legend.

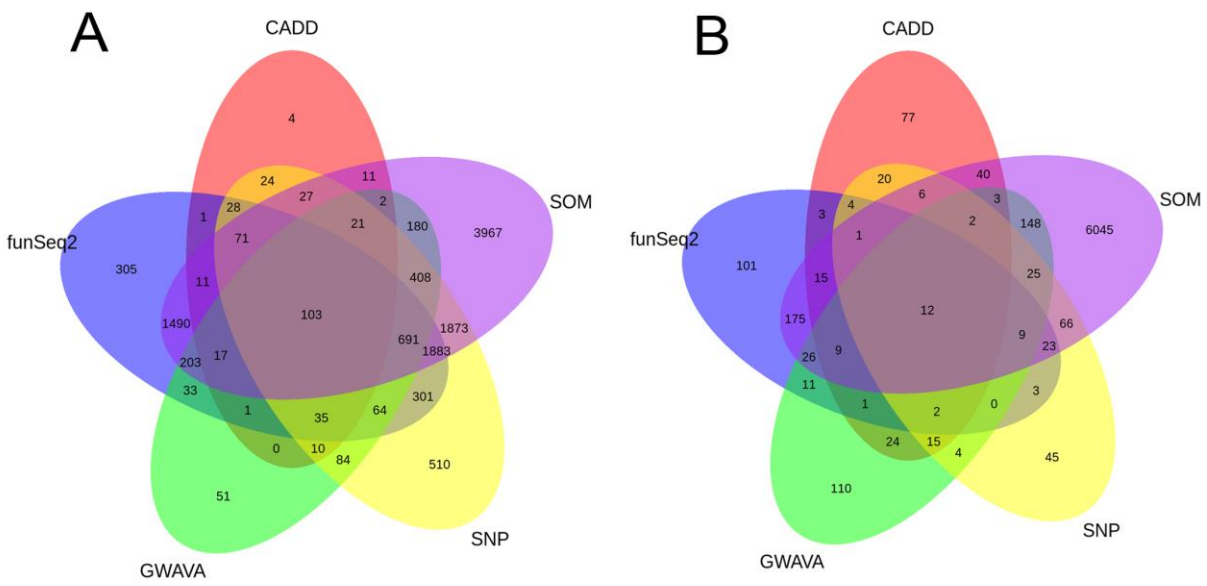

Figure S11. The comparison of driver candidates positively selected by five scoring tools in liver canccer. A. The overlap of the driver gene candidates predicted by the 5 permutation models (CADD, FunSeq2, GWAVA, SNP and SOM). B. The overlap of the lncRNA driver candidates predicted by the 5 permutation models (CADD, FunSeq2, GWAVA, SNP and SOM)

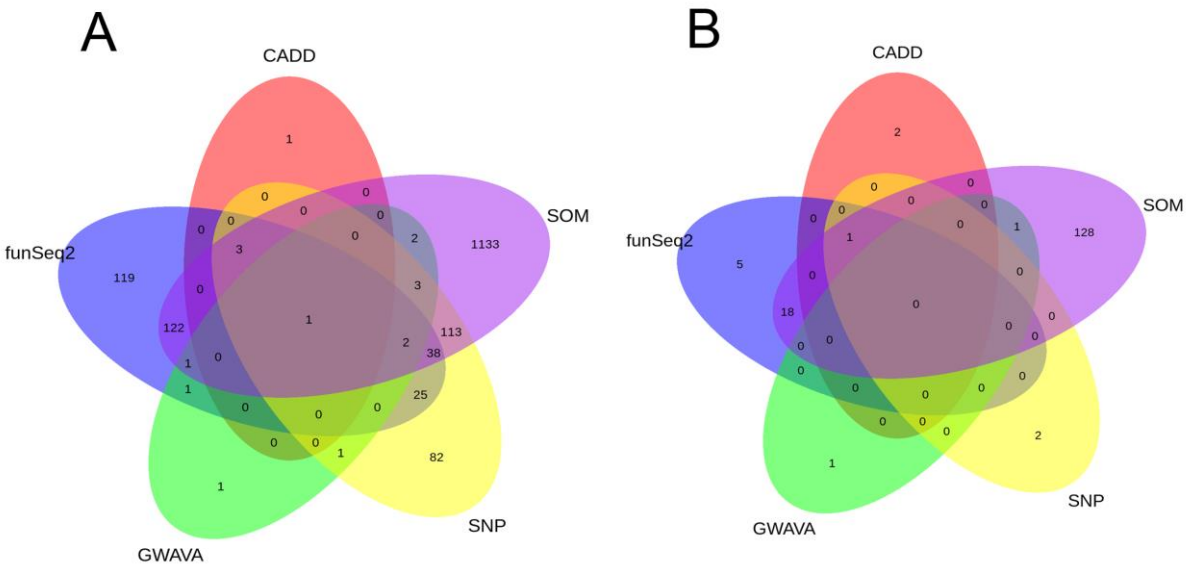

Figure S12. The comparison of driver candidates positively selected by five permutation models in CLL. See S11 for legend.

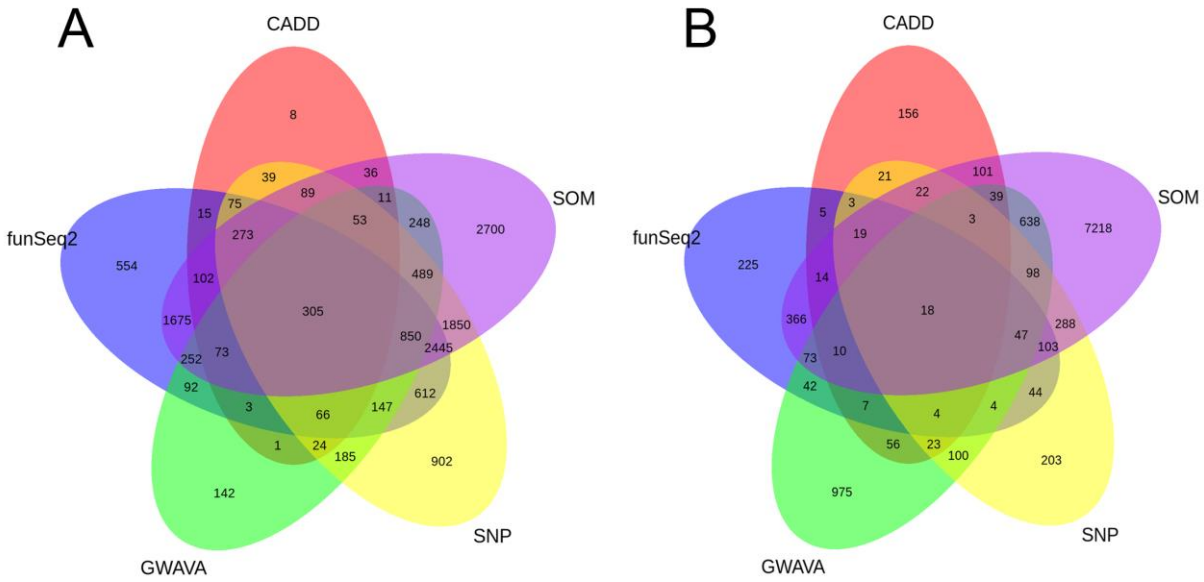

Figure S13. The comparison of driver candidates positively selected by five permutation models in melanoma. See S11 for legend.

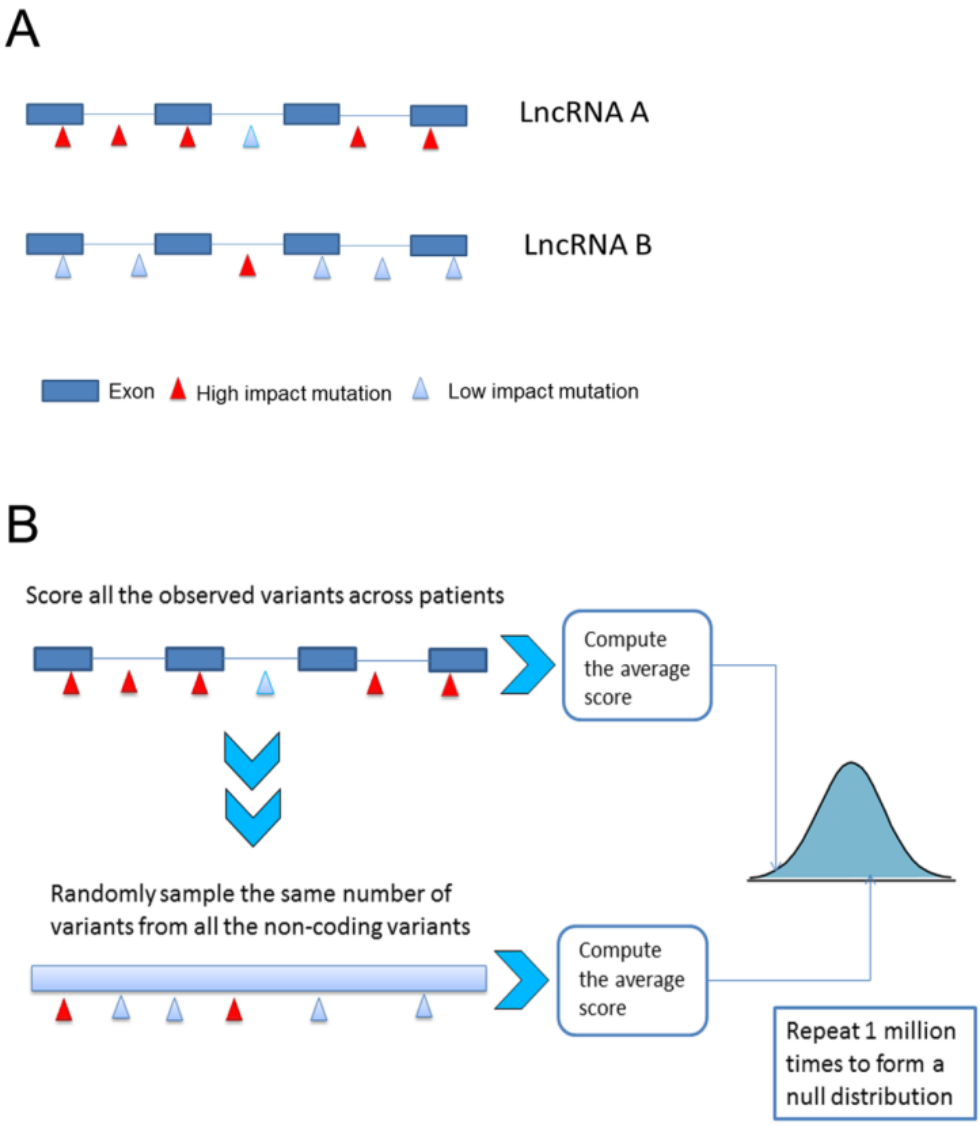

Figure S14. Detection of lncRNAs under positive selection in cancer

A. LncRNA A shows a higher enrichment of non-coding mutations with high function impact as compared to LncRNA B, indicating LncRNA A is under positive selection in cancer

B. The graphical display of permutation-based model for identifying lncRNAs harboring non-coding mutations with high function impact.

6.2 Supplemental Tables

Table S1. Uniform genomic features used in figures and SNP or SOM models.

Name	Description	Extent (Mb)	Reference	Model
UTR	mRNA untranslated region	47.23	Gencode v7(Harrow et al., 2012)	SNP+SOM
CDS	Coding sequence	35.34	Gencode	SOM
Exon.P	Exon of protein coding gene	91.15	Gencode	-
Intron.P	Intron of protein coding gene	1236.20	Gencode	SNP+SOM
PCgene	Protein coding gene	1266.97	Gencode	SOM
lncRNA	Long non-coding RNA	337.12	Gencode	SOM
Exon.L	Exon of lncRNA	16.44	Gencode	-
Intron.L	Intron of lncRNA	324.18	Gencode	SNP+SOM
ncExon	Non coding exon	30.61	Gencode	SNP+SOM
Intergenic	Intergenic region	1568.79	Gencode	SOM
5'SS	5'splicing site (10bp from the splicing site)	2.95	Gencode	-
3'SS	3'splicing site (50bp from the splicing site)	13.03	Gencode	-
GC content	Fraction of G or C nucleotide per 1Mb window	-	UCSC (Karolchik et al., 2014)	SOM
GC H	1-kb windows with high GC content (GC% > 50)	308.86	UCSC	-
GC L	1-kb windows with low GC content (GC% < 30)	104.89	UCSC	-
Promoter	Promoter	84.91	Gencode	SNP+SOM
Enhancer	Enhancer	12.03	FANTOM5(Andersson et al., 2014)	SNP
TFBS	Transcription factor binding site	947	ENCODE(Rosenbloom et al., 2013)	SNP
cTFBS	Conserved transcription factor binding site	59.23	UCSC	SNP+SOM
Sensitive	Khurana et al.'s region of high rate of rare SNP	9.21	(Khurana et al., 2013)	SOM
CR	Conserved region (PhastCons 46 way)	150.98	UCSC	SNP+SOM
ECS	Evolutionarily conserved RNA structure	199.68	(Smith et al., 2013)	SNP+SOM
DNase I	DNase I hypersensitive site (any cell type)	388.42	ENCODE	SNP+SOM
HE	Highly expressed gene/RNA (RPKM>20) in either cell line	635.78	ENCODE	SNP
LE	Low expressed gene/RNA (RPKM<0.25) in either cell line	1002.47	ENCODE	SNP
ER	Early replicated gene/RNA (EL ratio >1) in all cell lines	418.68	ENCODE	SNP
Recombination rate	Recombination rate averaged per 1Mb window	-	HAPMAP (Altshuler et al., 2010)	SOM
RR H	1-kb windows with high recombination rate (> 4.0)	117.55	HAPMAP	SNP
RR L	1-kb windows with low recombination rate (< 0.5)	1034.26	HAPMAP	SNP
GC	G or C base for each nucleotide	-	UCSC	SNP

Table S2. Cell-specific genomic features used in figures and SOM models.

Name	Description	Extent (Mb) of feature			
		Hepg2 (liver)	A549 (lung)	K562 (CLL)	Nhdfad (melanoma)
H3k4me1	H3k4me1	384.12	420.25	325.92	378.14
H3k4me2	H3k4m2	174.18	203.29	135.32	228.49
H3k4me3	H3k4me3	106.66	152.19	147.31	192.53
H3k9ac	H3k9ac	158.06	157.44	185.10	251.10
H3k9me3	H3k9me3	559.11	942.29	924.53	834.50
H3k27ac	H3k27ac	130.21	174.38	146.05	353.92
H3k27me3	H3k27me3	767.11	861.39	641.29	695.29
H3k36me3	H3k36me3	511.89	705.49	499.00	611.44
H3k79me2	H3k79me2	314.30	430.61	269.26	354.03
H3K20me1	H3K20me1	605.41	753.80	772.78	499.01
H2az	H2az	886.95	503.30	341.67	454.64
CTCF	CTCF	77.77	118.31	127.48	98.23
Ezh2	Ezh2	698.22	-	871.32	435.09
TFBS	Transcription factor binding site	286.38	164.61	348.46	65.69
Expression level	RPKM per 1Mb window	-	-	-	-
PCgene.HE	Highly expressed protein coding gene (RPKM >20)	93.08	72.36	101.08	79.21
PCgene.LE	Low expressed protein coding gene (RPKM <0.25)	422.35	222.52	457.92	311.64
LncRNA.HE	Highly expressed lncRNA (RPKM >20)	21.34	23.70	22.49	22.08
LncRNA.LE	Low expressed lncRNA (RPKM <0.25)	165.66	91.92	176.40	125.67
		Hepg2	Imr90	K562	Bg02
Replication time	Replication timing ratio per 1Mb window	-	-	-	-
LncRNA.early	Early replicated lncRNA (E/L ratio >1)	818.79	733.59	758.60	790.78
LncRNA.late	Late replicated lncRNA (E/L ratio <1)	441.3	520.16	497.73	471.30
PCgene.late	Late replicated protein coding gene (E/L ratio >1)	140.59	142.04	132.01	125.22
PCgene.early	Early replicated protein coding gene (E/L ratio <1)	182.39	175.87	188.71	198.26

		Liver hepatocelluar carcinoma	Lung adenocarcinoma	Acute myeloid leukemia	Skin cutaneous melanoma
DNA.met H	Average DNA methylation value < 0.4062	58.22	63.96	99.81	68.14
DNA.met L	Average DNA methylation value > 0.7245	58.22	57.26	51.85	52.38

Table S3. Significance of disease mutation enrichment in high-SNP+low SOM regions, for 4 cancer types.

Cancer type	Region	Region size (nt)	HGMD	Clivariant	P value (1)
-	Intergenic	1568807082	913	213	
-	High SNP	98163148	6784	1767	
Liver	Low SOM	1255672000	9719	4572	
	Low SOM+ high SNP	56198409	5079	1393	<2.2e-16
Lung	Low SOM	1186445000	9714	4596	
	Low SOM+ high SNP	56160584	5012	1391	<2.2e-16
CLL	Low SOM	1236512000	9580	4660	
	Low SOM+ high SNP	56267795	4773	1332	<2.2e-16
Melanoma	Low SOM	1170977000	9265	4384	
	Low SOM+ high SNP	56148149	4892	1322	<2.2e-16

(1) P values are computed as follows: disease-associated variants from the HGMD or Clivariant database are counted in high SNP or low SOM *vs*. Low SOM+high SNP regions, along with region sizes, forming a 2x2 matrix for Chi-square test in each cancer type. P values here are statistical significance for both HGMD and Clivariant databases.

Table S4. Significance of over-enrichment for hypomutated regions within cancer vs non-cancer genes. Enrichment for hypomutated regions was computed as explained in Methods for each independent gene. Then for each gene class (protein-coding, lncRNA, miRNA), a Wilcoxon rank sum test was performed to compare enrichment factors in cancer genes (see

Methods for gene lists) and in all genes in the class.

Gene type	Cancer type	P-value
Protein-coding (non-coding parts)	Liver	1.44E-12
	Lung	5.05E-14
	CLL	5.22E-15
	Melanoma	<2.20E-16
lncRNA	Liver	0.838
	Lung	0.158
	CLL	0.705
	Melanoma	0.903
miRNA	Liver	0.007
	Lung	0.003
	CLL	0.011
	Melanoma	0.004

Table S5: Biological process GO-term biases (1) in the 100 protein coding genes with highest coverage by hypermutated (high SNP-high SOM) positions (liver cancer and CLL).

GO biological process complete	**#**	**#**	**expected**	**Fold Enrichment**	**+/-**	**P value (2)**
Liver cancer						
Unclassified	4272	17	18.88	.90	-	0.00E00
transcription from RNA polymerase II promoter	781	19	3.45	> 5	+	1.04E-05
gene expression	3825	41	16.91	2.43	+	5.55E-05
cellular nitrogen compound metabolic process	5112	48	22.60	2.12	+	9.12E-05
nucleobase-containing compound metabolic process	4372	43	19.32	2.23	+	2.58E-04
RNA metabolic process	3373	37	14.91	2.48	+	2.61E-04

nucleic acid metabolic process	3874	40	17.12	2.34	+	2.85E-04
cellular nitrogen compound biosynthetic process	3407	37	15.06	2.46	+	3.42E-04
nucleobase-containing compound biosynthetic process	2962	34	13.09	2.60	+	4.21E-04
RNA biosynthetic process	2680	32	11.85	2.70	+	5.01E-04
transcription, DNA-templated	2560	31	11.32	2.74	+	6.41E-04
nucleic acid-templated transcription	2561	31	11.32	2.74	+	6.47E-04
heterocycle biosynthetic process	3043	34	13.45	2.53	+	8.19E-04
aromatic compound biosynthetic process	3044	34	13.45	2.53	+	8.25E-04
nitrogen compound metabolic process	5475	48	24.20	1.98	+	9.12E-04
Chronic lymphocytic leukemia (CLL)						
positive regulation of transcription from RNA polymerase II promoter	987	17	4.46	3.81	+	1.52E-02

Table S6. Mammalian long non-coding RNAs experimentally shown to be associated with different cancer types from a literature search.

Chromosome	Start	End	LncRNA	Size(bp)	Reference
chr9	21994789	22029563	ANRIL	503	(Kotake et al., 2011)
chr1	173833038	173837125	GAS5	632	(M. Sun et al., 2014)
chr12	54356095	54362515	HOTAIR	2337	(Gupta et al., 2010)
chr7	27135712	27139585	HOTAIRM1	483	(X. Zhang et al., 2014)
chr6	8652441	8654459	HULC	500	(Panzitt et al., 2007)
chr3	116428634	116435887	LOC285194	2105	(Qian Liu et al., 2013)
chr3	50137035	50138421	LUST	1386	(Rintala-Maki and Sutherland, 2009)
chr6	136265388	136282959	NTT	17572	(Delgado Andr é and De Lucca, 2008)
chr9	79379353	79402465	PCA3	3735	(Gezer et al., 2015)
chr8	128025398	128033259	PCAT1	1992	(Prensner et al., 2011)
chr2	193614570	193641625	PCGEM1	1590	(L. Yang et al., 2013)
chr9	33673501	33677418	PTENP1	3932	(C.-L. Chen et al., 2015)

chr3	181417385	181433076	Sox2ot	2970	(Askarian-Amiri et al., 2014)
chr5	139929652	139937678	SRA	1965	(Leygue et al., 1999)
chr22	31365633	31375381	TUG1	7105	(E. Zhang et al., 2014)
chr19	15939756	15946230	UCA1	1413	(C. Yang et al., 2012)
chrX	73040494	73072588	XIST	19271	(McHugh et al., 2015)
chr8	128092118	128104845	PRNCR1	12756	(L. Yang et al., 2013)
chr14	101292444	101327363	MEG3	1855	(Benetatos et al., 2011)
chr11	2016405	2019065	H19	2308	(Fellig et al., 2005)
chr11	65265232	65273940	MALAT1	8708	(Ji et al., 2003)
chr14	61283510	61285560	HIF1A-AS2	2050	(W. Chen et al., 2015)
chr17	23111183	23134213	Anti-NOS2A	23	(Korneev et al., 2008)
chr7	148315552	148317449	GHET1	1898	(F. Yang et al., 2014)
chr20	5048232	5048615	PCNA-AS1	384	(Yuan et al., 2014)

Table S7. Significance of adjusted CADD, combined, funSeq2, GWAVA, SNP and SOM P values between cancer genes and protein coding genes, cancer lncRNAs and lncRNAs for 4 cancer types.

Comparison	Cancer type	CADD	Combined	FunSeq2	GWAVA	SNP	SOM
Cancer gene VS PCgene	Liver cancer	1,77E-039	4,91E-034	1,70E-089	3,69E-030	2,61E-029	1,05E-012
	Lung cancer	2,23E-044	1,43E-027	4,97E-094	2,60E-035	1,50E-022	6,98E-011
	CLL	3,17E-015	3,46E-016	9,09E-034	6,14E-021	1,21E-008	5,46E-005
	Melanoma	4,62E-044	3,48E-032	7,73E-092	6,70E-032	7,16E-025	2,95E-013
Cancer lncRNA VS lncRNA	Liver cancer	1,07E-004	1,97E-008	4,74E-037	1,84E-019	2,21E-042	2,13E-003
	Lung cancer	1,91E-004	8,84E-006	2,17E-019	3,66E-013	1,20E-007	4,83E-004
	CLL	1,34E-001	3,34E-004	3,05E-049	1,72E-016	1,37E-030	4,69E-004
	Melanoma	4,40E-002	5,09E-004	5,39E-004	1,29E-003	2,38E-019	2,30E-002

P values are computed as follows: adjusted CADD, combined, funSeq2, GWAVA, SNP and SOM P values were compared with Wilcoxon rank sum test between cancer gene and protein coding genes, cancer lncRNAs and lncRNAs successively for each cancer type.

Table S8. Adjusted P values and P value rankings of top 10 recurrently mutated genes in liver cancer

RMG	CADD	FunSeq2	GWAVA	SNP	SOM	Combined
	Adjusted P value (Ranking of P value)					
TP53	0,0000(1)	0,0000(1)	1,0000(2007)	0,0013(257)	0,0000(1)	0,0000(1)
CTNNB1	0,6944(2748)	0,0000(1)	0,1135(857)	0,0001(10)	0,0001(40)	0,0000(1)
TERT	1,0000(4953)	0,0786(2714)	1,0000(2007)	0,7601(6498)	0,0000(1)	0,0000(1)
ARID1A	0,0644(229)	0,0000(1)	0,0000(1)	0,0000(1)	0,0000(1)	0,0000(1)
HNF1A	0,6144(2150)	0,0070(884)	0,6690(1635)	0,2201(4272)	0,0000(3)	0,0000(1)
AXIN1	0,3937(1106)	0,0435(2140)	1,0000(2007)	0,0144(1497)	0,0001(24)	0,0031(1866)
ARID2	0,6259(2239)	0,1830(3668)	0,0000(1)	0,0000(1)	0,0000(1)	0,0000(1)
IL6ST	0,7880(3435)	0,0000(2)	0,0520(637)	0,0097(1178)	0,0000(8)	0,0000(1)
CDKN2A	0,6653(2543)	0,0001(13)	0,0778(739)	0,4164(5333)	0,3600(4717)	0,0005(1)
ATM	0,3417(940)	0,0000(1)	0,0005(15)	0,0006(123)	0,0001(27)	0,0000(1)
Number of unique P values	4953	6632	2007	7300	5406	7496

Table S9. Adjusted P values and P value rankings of top 10 recurrently mutated genes in CLL

RMG	CADD	FunSeq2	GWAVA	SNP	SOM	Combined
	Adjusted Pvalue (Ranking of P value)					
MED12	0,9064(466)	0,2413(553)	0,3710(73)	0,0489(138)	0,0784(1038)	0,0049(457)
POT1	0,7223(180)	0,1020(308)	1,0000(190)	0,3209(667)	0,5267(1825)	0,2263(1343)
BCL11B	0,6437(112)	0,1570(421)	0,1986(37)	0,1191(313)	0,0927(1133)	0,0004(220)
EGFR	0,8260(335)	0,0023(26)	0,8373(155)	0,4373(837)	0,0265(551)	0,0033(400)
BCOR	0,7224(181)	0,2678(584)	0,8505(157)	0,0702(198)	0,1142(1231)	0,0100(558)
ROS1	0,7468(207)	0,2714(591)	1,0000(190)	0,5062(923)	0,1796(1468)	0,2328(1358)
Number of unique P values	627	1118	190	1279	2003	1974

Table S10. Adjusted P values and P value rankings of top 10 recurrently mutated genes in melanoma

RMG	CADD	FunSeq2	GWAVA	SNP	SOM	Combined
	Adjusted Pvalue (Ranking of P value)					
BRAF	0,0539(641)	0,0000(1)	0,0055(286)	0,0000(1)	0,0000(1)	0,0000(1)
TP53	0,2796(2179)	0,0001(25)	0,3713(1756)	0,0022(584)	0,0000(1)	0,0000(1)
ARID2	0,2101(1635)	0,0805(3337)	0,0003(17)	0,0000(1)	0,0000(1)	0,0000(1)
ARID1A	0,0281(415)	0,0000(1)	0,0001(3)	0,0000(1)	0,0000(1)	0,0000(1)
MAP2K1	0,5142(4061)	0,0000(1)	0,7810(2301)	0,0000(1)	0,0000(1)	0,0000(1)
FGFR3	0,3749(2864)	0,0000(2)	1,0000(2575)	1,0000(7616)	0,0001(30)	0,0038(3224)
BCL9	0,2414(1862)	0,1017(3587)	0,4900(1925)	0,0000(1)	0,0001(23)	0,0000(1)
NCOA1	0,3326(2568)	0,0000(1)	0,0006(34)	0,0000(1)	0,0000(1)	0,0000(1)
PAX3	0,0216(348)	0,0000(1)	1,0000(2575)	0,0038(846)	0,0049(1288)	0,0000(1)
TPM3	0,1027(986)	0,0000(1)	0,0012(82)	0,0000(1)	0,0000(1)	0,0000(1)
Number of unique P values	7752	7409	2575	7616	6482	7218

Table S11.the driver candidates of PCgenes and lncRNAs positively selected by each model in liver cancer

Tool	Adjusted P values < 0.05 Number of genes (Mb)		Average length (bp)	
	PCgene	LncRNA	PCgene	LncRNA
CADD	366(103)	234(19)	283679	83942
funSeq2	5237(589)	395(21)	112575	55171
GWAVA	1903(214)	401(23)	112642	58917
SNP	6133(780)	237(22)	127337	95182
SOM	10958(880)	6605(187)	80390	28412
Combined	9739(867)	2821(113)	89113	40170
Total Genes	20300(1266)	38263(456)	62412	11917

Table S12.the driver candidates of PCgenes and lncRNAs positively selected by each model in CLL

Tool	Adjusted P values < 0.05 Number of genes (Mb)		Average length (bp)	
	PCgene	LncRNA	PCgene	LncRNA
CADD	5(1)	3(0.54)	210741	180329
funSeq2	312(130)	24(1)	416712	47786
GWAVA	12(4)	2(0.14)	337753	69675
SNP	268(115)	3(0.33)	432029	109058
SOM	1418(330)	148(22)	232755	154102
Combined	1144(307)	94(113)	268402	106835
Total Genes	20300(1266)	38263(456)	62412	11917

Table S13.the driver candidates of PCgenes and lncRNAs positively selected by each model in melanoma

Tool	Adjusted P values < 0.05 Number of genes (Mb)		Average length (bp)	
	PCgene	LncRNA	PCgene	LncRNA
CADD	1173(238)	501(34)	202996	69450
funSeq2	7539(701)	984(35)	93109	35929
GWAVA	2491(282)	2137(59)	96088	27912
SNP	8404(942)	1000(52)	112206	52286
SOM	11451(883)	9057(195)	77119	21606
Combined	11322(852)	5066(129)	75323	25611
Total Genes	20300(1266)	38263(456)	62412	11917

Table S14.LncRNA driver candidates common to five permutation models in lung cancer

Chromosome	Start	End	LncRNA	CADD	FunSeq2	GWAVA	SNP	SOM	Combined
chr17	46667781	46683774	HOXB-AS3	0,0325	0,0001	0,0072	0,0093	0,0001	0,0000
chr1	245003940	245018799	HNRNPU-AS1	0,0412	0,0012	0,0000	0,0000	0,0006	0,0000
chr11	57479994	57586652	TMX2-CTNND1	0,0279	0,0000	0,0004	0,0000	0,0000	0,0000
chr2	179385910	179639402	TTN-AS1	0,0000	0,0000	0,0000	0,0069	0,0000	0,0000
chr2	144433734	144498863	RP11-434H14.1	0,0000	0,0000	0,0000	0,0000	0,0109	0,0000
chr7	27147396	27173921	HOXA-AS2	0,0103	0,0000	0,0068	0,0120	0,0000	0,0000
chr12	54747474	54860814	LOC102724050	0,0144	0,0000	0,0000	0,0006	0,0000	0,0000
chr12	54747576	54860769	RP11-753H16.5	0,0144	0,0000	0,0000	0,0006	0,0000	0,0000

chr2	179246804	179541009	MIR548N	0,0000	0,0000	0,0000	0,0036	0,0000	0,0000
chr2	144052990	144238358	AC096558.1	0,0000	0,0000	0,0000	0,0000	0,0060	0,0000
chr2	144053155	144329674	RP11-570L15.2	0,0000	0,0000	0,0000	0,0000	0,0002	0,0000

Table S15.LncRNA driver candidates common to five permutation models in liver cancer

Chromosome	Start	End	LncRNA	CADD	FunSeq2	GWAVA	SNP	SOM	Combined
chr17	37558046	37562486	CTB-131K11.1	0,0361	0,0208	0,0281	0,0335	0,0099	0,0000
chr1	155996957	156132001	MIR7851	0,0099	0,0000	0,0058	0,0010	0,0000	0,0000
chr11	57479994	57586652	TMX2-CTNND1	0,0160	0,0125	0,0111	0,0310	0,0000	0,0000
chr20	39726969	39766643	RP1-1J6.2	0,0442	0,0000	0,0000	0,0000	0,0000	0,0000
chr20	39726633	39766640	PLCG1-AS1	0,0442	0,0000	0,0000	0,0000	0,0000	0,0000
chr3	114172440	114238979	RP11-197K3.1	0,0017	0,0000	0,0000	0,0002	0,0056	0,0000
chr3	114172439	114238979	LOC101929754	0,0017	0,0000	0,0000	0,0002	0,0056	0,0000
chr3	99273152	99717059	MIR548G	0,0043	0,0008	0,0000	0,0000	0,0000	0,0000
chr12	11944823	12079107	RNU6-19P	0,0256	0,0060	0,0014	0,0002	0,0000	0,0000
chr2	179246804	179541009	MIR548N	0,0000	0,0000	0,0000	0,0459	0,0000	0,0000
chr15	60771377	60922836	RP11-219B17.1	0,0014	0,0000	0,0000	0,0004	0,0000	0,0000

Table S16.LncRNA driver candidates common to five permutation models in melanoma

Chromosome	Start	End	LncRNA	CADD	FunSeq2	GWAVA	SNP	SOM	Combined
chr17	56402811	56493127	BZRAP1-AS1	0,0049	0,0000	0,0000	0,0002	0,0000	0,0000
chr12	54656399	54672847	RP11-968A15.2	0,0181	0,0001	0,0019	0,0073	0,0177	0,0000
chr7	27147396	27173921	HOXA-AS2	0,0000	0,0000	0,0000	0,0005	0,0000	0,0000
chr20	39726969	39766643	RP1-1J6.2	0,0486	0,0000	0,0083	0,0000	0,0002	0,0000
chr20	39726633	39766640	PLCG1-AS1	0,0486	0,0000	0,0083	0,0000	0,0002	0,0000
chr15	72571208	72644135	RP11-106M3.3	0,0438	0,0117	0,0101	0,0463	0,0000	0,0000
chr12	54670415	54738867	RP11-968A15.8	0,0124	0,0000	0,0000	0,0000	0,0000	0,0000
chr1	243866159	243904123	RP11-370K11.1	0,0057	0,0000	0,0009	0,0001	0,0000	0,0000
chr1	33452676	33498070	RP1-117O3.2	0,0289	0,0005	0,0312	0,0000	0,0000	0,0000
chr17	41622153	41687706	RP11-392O1.4	0,0248	0,0000	0,0017	0,0000	0,0000	0,0000
chr12	54747474	54860814	LOC102724050	0,0000	0,0000	0,0002	0,0032	0,0000	0,0000
chr12	54747576	54860769	RP11-753H16.5	0,0000	0,0000	0,0002	0,0032	0,0000	0,0000
chr2	144052990	144238358	AC096558.1	0,0000	0,0000	0,0026	0,0001	0,0000	0,0000
chr1	23346640	23414551	RP1-184J9.2	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000
chr14	30637040	30766245	TCONS_00022407	0,0000	0,0009	0,0088	0,0167	0,0000	0,0000
chr2	179385910	179639402	TTN-AS1	0,0440	0,0000	0,0006	0,0271	0,0010	0,0000
chr15	60771377	60922836	RP11-219B17.1	0,0000	0,0000	0,0000	0,0046	0,0072	0,0000
chr3	99273152	99717059	MIR548G	0,0020	0,0000	0,0000	0,0000	0,0000	0,0000

Table S17. Significance of overlap of CAAD, funSeq2, GWAVA, SNP and SOM driver candidates in 4 cancer types.

Gene	Liver cancer	Lung cancer	CLL	Melanoma
PC gene	0	0	0	0
LncRNA	0	0	1	0

The significance of overlap was computed for CAAD, funSeq2, GWAVA, SNP and SOM driver candidates in 4 cancer types using a permutation test as follows: the same number of protein coding genes or lncRNAs with the driver candidates was randomly sampled from the whole lncRNAs set 1000 times, the overlap was calculated for five random sampling genes or lncRNAs. Then a P value was generated via comparing the observed overlap of driver candidates with 1000 sampling ones.

Table S18. Significance of enrichment of conserved regions for CAAD, combined, funSeq2, GWAVA, SNP and SOM lncRNA driver candidates in 4 cancer types.

Cancer type	CADD	Combined	FunSeq2	GWAVA	SNP	SOM
Liver cancer	0	0	0	0	0	0.911
Lung cancer	0	0	0	0	0	0.057
CLL	0,07	0	0.003	0.446	0.009	0.08
Melanoma	0	0	0	0	0	0.028

The enrichment of conserved regions was calculated as described in the method. The significance of enrichment of conserved regions was computed using a permutation test as follows: a set of positions of same size as the driver candidate (ie. 17.31 Mb) was randomly sampled from the whole lncRNAs set 1000 times, and in each random sample, enrichment was calculated for each driver candidate class. Then a P value was generated via comparing the observed enrichment of conserved regions with 1000 sampling ones.

Table S19. Significance of enrichment of HGMD and Clivariant disease-causing variants for CADD, combined, funSeq2, GWAVA, SNP and SOM lncRNA driver candidates in 4 cancer types.

Disease mutation	Cancer type	CADD	Combined	FunSeq2	GWAVA	SNP	SOM
HGMD	Liver cancer	0	0	0	0	0	0
	Lung cancer	0.434	0	0.011	0.483	0.035	0.201
	CLL	0.14	1	0.062	0.114	0.26	0.154
	Melanoma	0.565	0.002	0	0.235	0.108	0.11
Clivariant	Liver cancer	0	0	0	0	0	0
	Lung cancer	0.032	0.025	0.007	0.023	0.014	0.338
	CLL	0.211	0.151	0.418	0.016	0.148	0.962
	Melanoma	0.115	0.007	0	0.009	0.016	0

The enrichment of HGMD and Clivariant disease-causing variants was calculated as described in the method, the significance of enrichment of disease variants was computed using the same permutation test as S11 (See S19 for method).

Table S20. 61 Mammalian long non-coding RNAs experimentally shown to be associated with different cancer types from a literature search.

Chromosome	Start	End	LncRNA	Length(bp)
chr14	101292444	101327363	MEG3	34919
chr11	2016405	2019065	H19	2660
chr11	65265232	65273940	MALAT1	8708
chr14	61283510	61285560	HIF1A-AS2	2050
chr17	23111183	23134213	NOS2A	23030

chr7	148315552	148317449	GHET1	1897
chr9	21994789	22029563	ANRIL	34774
chr1	173833038	173837125	GAS5	4087
chr12	54356095	54362515	HOTAIR	6420
chr7	27135712	27139585	HOTAIRM1	3873
chr6	8652441	8654459	HULC	2018
chr3	50137035	50138421	LUST	1386
chr6	136265388	136282959	NTT	17571
chr2	192749845	192776899	PCGEM1	27054
chr5	139929652	139937678	SRA	8026
chr22	31365633	31375381	TUG1	9748
chrX	73040494	73072588	XIST	32094
chr15	69463026	69571440	RP11-279F6.1	108414
chr8	126847055	127021014	PCAT1	173959
chr7	77657660	77697345	APTR	39685
chr3	180989770	181836880	SOX2-OT	847110
chr20	5119586	5119969	PCNA-AS1	383
chr8	127079874	127092595	PRNCR1	12721
chr21	36131767	36175815	PlncRNA-1	44048
chr17	76557764	76565348	ncRAN	7584
chr3	116921431	116932238	BC040587	10807
chr9	33673504	33677499	PTENP1	3995
chr7	27198575	27207259	HOTTIP	8684
chr1	202810954	202812156	PCAN-R1	1202
chr9	94555069	94568127	PCAN-R2	13058
chr9	69296681	69307056	BANCR	10375
chr19	15828947	15836321	UCA1	7374
chr16	74701404	74702604	lncRNA-EBIC	1200
chr17	42865922	42874369	AOC4P	8447

chr10	31206278	31320447	ZEB1-AS1	114169
chr14	19858667	19941024	lnc-ATB	82357
chr12	120941728	120980965	HNF1A-AS1	39237
chr15	69463026	69571440	DRAIC	108414
chr15	69592129	69695750	PCAT29	103621
chr7	27193503	27200106	HOXA13	6603
chr20	50040707	50041629	treRNA	922
chr8	75223404	75278461	ESCCAL-1	55057
chr20	56285239	56287836	NKILA	2597
chr19	15828947	15836321	CUDR	7374
chr6	36673621	36675126	PANDAR	1505
chr20	30309310	30311212	INXS	1902
chr10	4769152	4772545	uc002mbe.2	3393
chr1	168873143	169056243	AK126698	183100
chr18	57054559	57072119	lincRNA-RoR	17560
chr3	116921431	116932238	BC040587	10807
chr9	33673504	33677499	PTENP1	3995
chr7	27198575	27207259	HOTTIP	8684
chr1	202810954	202812156	PCAN-R1	1202
chr9	94555069	94568127	PCAN-R2	13058
chr9	69296681	69307056	BANCR	10375
chr19	15828947	15836321	UCA1	7374
chr16	74701404	74702604	lncRNA-EBIC	1200
chr17	42865922	42874369	AOC4P	8447
chr10	31206278	31320447	ZEB1-AS1	114169
chr14	19858667	19941024	lnc-ATB	82357
chr12	120941728	120980965	HNF1A-AS1	39237
chr15	69463026	69571440	DRAIC	108414
chr15	69592129	69695750	PCAT29	103621

chr7	27193503	27200106	HOXA13	6603
chr20	50040707	50041629	treRNA	922
chr8	75223404	75278461	ESCCAL-1	55057
chr20	56285239	56287836	NKILA	2597
chr19	15828947	15836321	CUDR	7374
chr6	36673621	36675126	PANDAR	1505
chr20	30309310	30311212	INXS	1902
chr10	4769152	4772545	uc002mbe.2	3393
chr1	168873143	169056243	AK126698	183100
chr18	57054559	57072119	lincRNA-RoR	17560
chr2	87455368	87606805	Linc00152	151437
chr5	180829954	180831618	lncRNA-HEIH	1664
chr9	76764436	76787569	DD3(PCA3)	23133
chr5	141697199	141697887	SPRY4-IT1	688
chr3	116709235	116723581	LOC285194	14346
chr8	127890589	128101253	PVT1	210664
chr5	140102922	140107643	MA-linc1	4721
chr16	53071943	53073640	PR-lncRNA-1	1697
chr9	139001797	139004427	PR-lncRNA-10	2630
chr6	36632321	36635073	lincRNA-p21	2752

6.3 Supplemental Methods

Random Forest Models

The random forests (RF) approach involves producing multiple regression trees, which are then combined to make a single consensus prediction for a given observation (Breiman L, 2001). We generated the SNP RF model and the SOM RF model using the *randomForest* R package. The RF model is composed of an aggregate collection of regression trees, each created from boostrapped training samples: each branch is selected from a random subset of a given number (denoted be *mtry*) of the input variables (data columns). The two main parameters are *mtry* and *ntree*, the number of trees in the forest. We used the mean squared error (abbreviated MSE) as a measure of the prediction accuracy of the RF model. Two MSE error estimates are used in the validation procedure: the OOB error and the cross-validation error. An important feature of RFs is its use of *out-of-bag* (OOB) samples. An OOB sample is the set of observations which are not used for building the current tree, and can be used to estimate the MSE error; it can be shown that an OOB error estimate is almost identical to that obtained by K-fold cross-validation.

RF models have the advantage of giving a summary of the importance of each variable based on the randomized variable selection process used to grow the RF. An estimation of variable importance is provided by *IncNodePurity*, which measures the decrease in tree node purity that results from all splits of a given variable over all trees. This measure can be used to rank variables by the strength of their relation to the response variable, for interpretation purposes.

Model Calibration

We first tuned the two parameters *mtry* and *ntree* of the RF method. Figure S15 shows the OOB error progression on 500 trees for random forests using different parameters *mtry*. MSE errors stabilize at about 400 trees, so we see that *ntree*=500 (default value) was sufficient to give good performance for the SNP model and for the SOM model.

In a regression framework, the default value of *mtry* is [p/3] where p is the number of variables. The case *mtry*=p corresponds to bagging (or bootstrap aggregation), a general purpose procedure for reducing the variance of a statistical learning method. For the SNP data

we have p=18 and the default value of mtry is 5. Note that a larger mtry is best suited to the SNP and SOM data, according to the MSE error (Figure S15 and Figure S18). We considered the gain in MSE error was small enough for mtry greater than 7 for the SNP model and 10 for the liver cancer SOM model.

Assessment of variable importance is performed using *IncNodePurity*, with larger values indicating more important variables. We examined the RF variable importances behavior for different values of *ntree* and *mtry*. In Figure S16 and Figure S19, a graphical representation of the variable importances is shown using 3 values of mtry (5 the default, 7 and 14 for SNP model, 10, 20 and 30 for SOM model of liver cancer) and two values of *ntree* (the 500 default and 1000). The magnitude of the variable importances is increased with larger values of *mtry*, but we get nearly the same order for all variables in every run of the procedure and with every value of *mtry*. Moreover, using a small value of mtry is preferred in the presence of correlated predictors. We chose mtry=7 for the SNP model and mtry=10 for the SOM models of liver cancer, lung cancer, CLL and melanoma, respectively, based on lower MSE errors and smaller mtry values (Figure S15 and Figure S18).

Feature selection

We used the R *VSURF* package to perform variable selection. The selection procedure is based on a ranking of the explanatory variables using the random forests score of importance and a stepwise ascending strategy (Genuer, 2010). The first step eliminates the noisy variables and the second step selects the variables leading to the smallest OOB error. One advantage in using the VSURF procedure lies in its robustness with respect to the choice of *mtry* and *ntree*.

Model Validation

RFs were grown with *ntree*=500 for all models. We used *mtry*=7 for the SNP model and mtry=10 for the SOM models. The SNP RF model was trained using 16 explanatory variables. The SOM RF models were estimated using 21, 22, 29 and 23 explanatory variables selected by VSURF for liver cancer, lung cancer, CLL and melanoma respectively. The validation of the two models is given in terms of MSE. We used 10-fold cross-validation to compute the prediction error. We compared the prediction error of the RF model to the prediction error obtained training a multiple regression linear model with the same input variables involved.

We see that RFs outperform a linear model for the SNP and cancer mutation data (Figure S17A and Figure S20A).

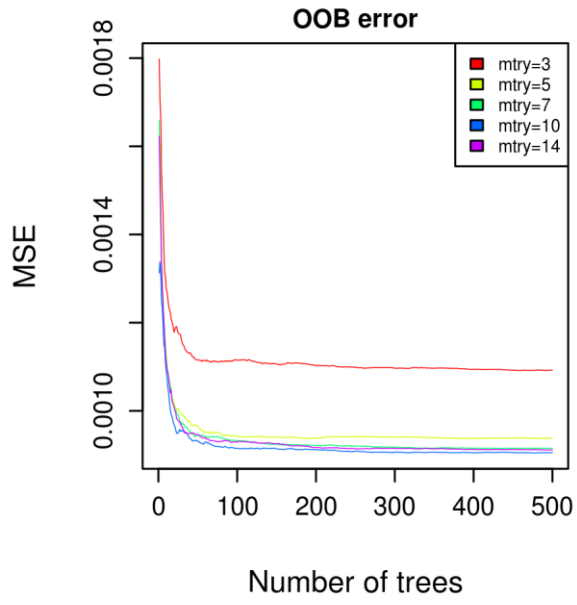

Figure S15. MSE sensitivity to ntree and mtry (SNP model)

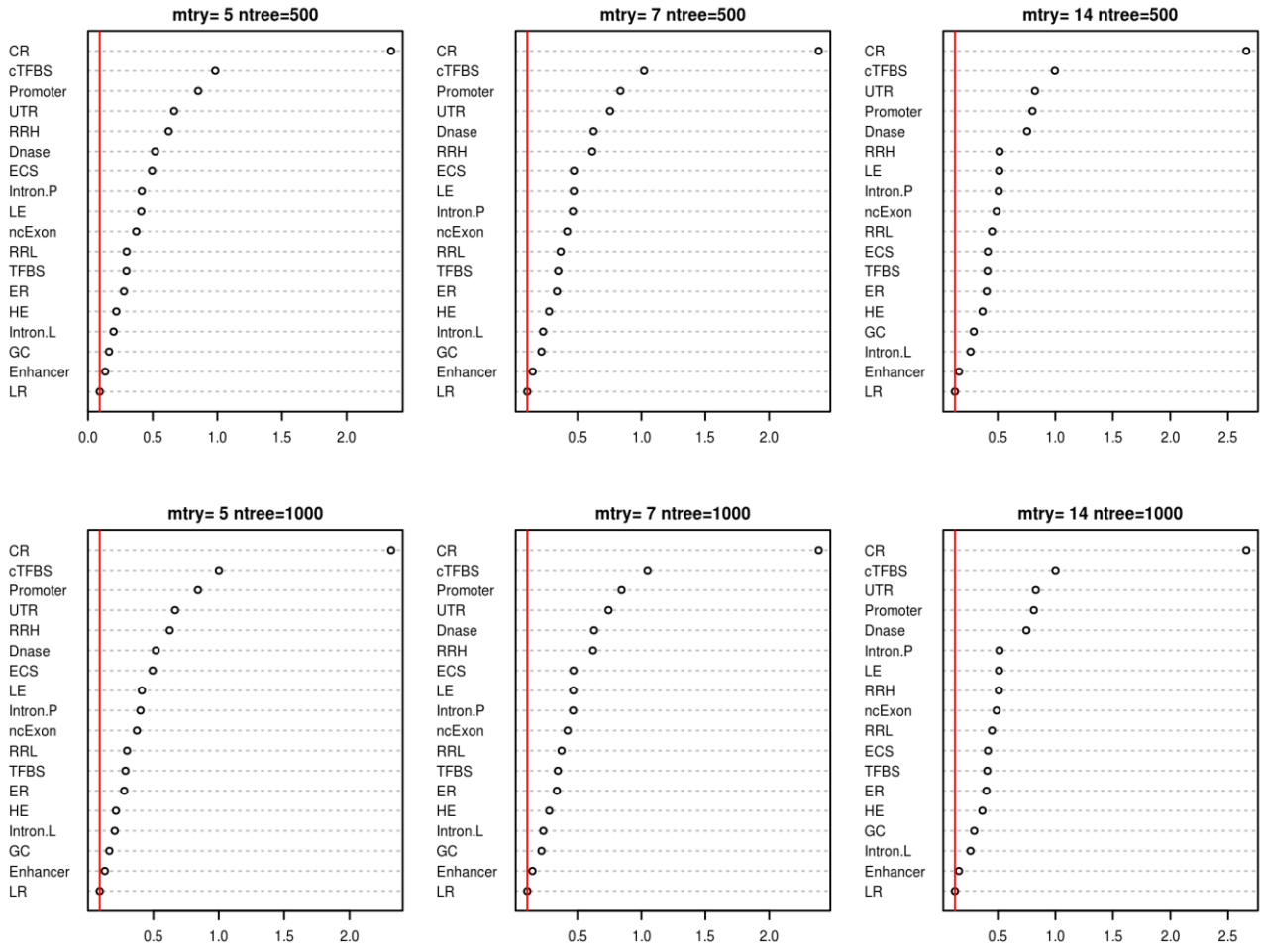

Figure S16. variable importance (lncNodePurity) sensitivity to ntree and mtry (red line: absolute value of minimum importance among all features in the SNP RF model)

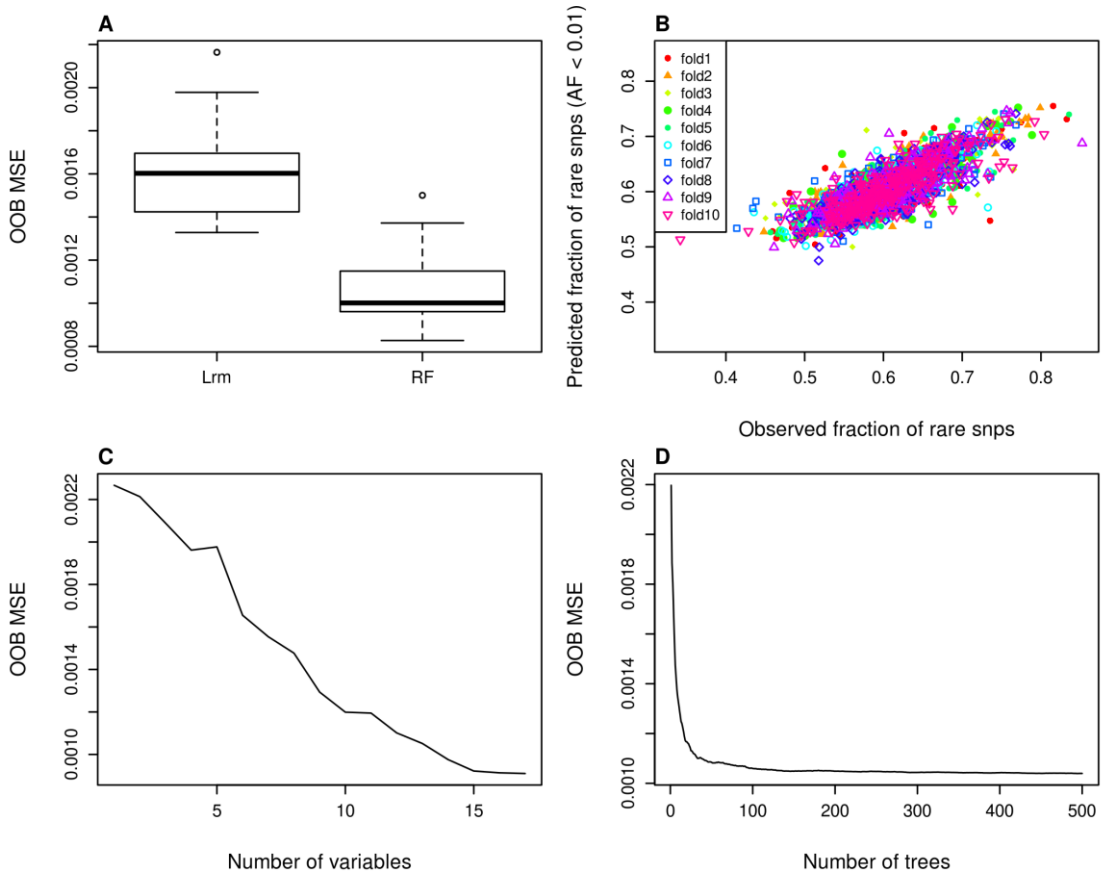

Figure S17. Validation of the SNP model. A. MSE for linear regression model (Lrm) and Random forest (RF) with 10-fold cross validation (SNP model); B. observed and predicted fraction of rare SNPs (AF < 0.01) with 10-fold cross validation; C. the number of variables remained in the RF model minimizes the OOB error; D. the default number of trees in the RF model minimizes the OOB error.

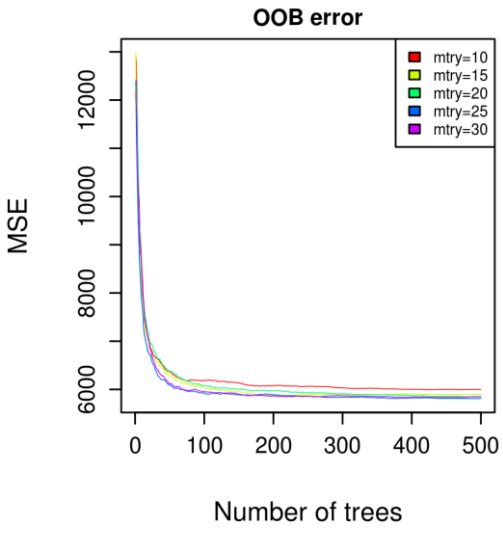

Figure S18. MSE sensitivity to ntree and mtry (SOM liver cancer model)

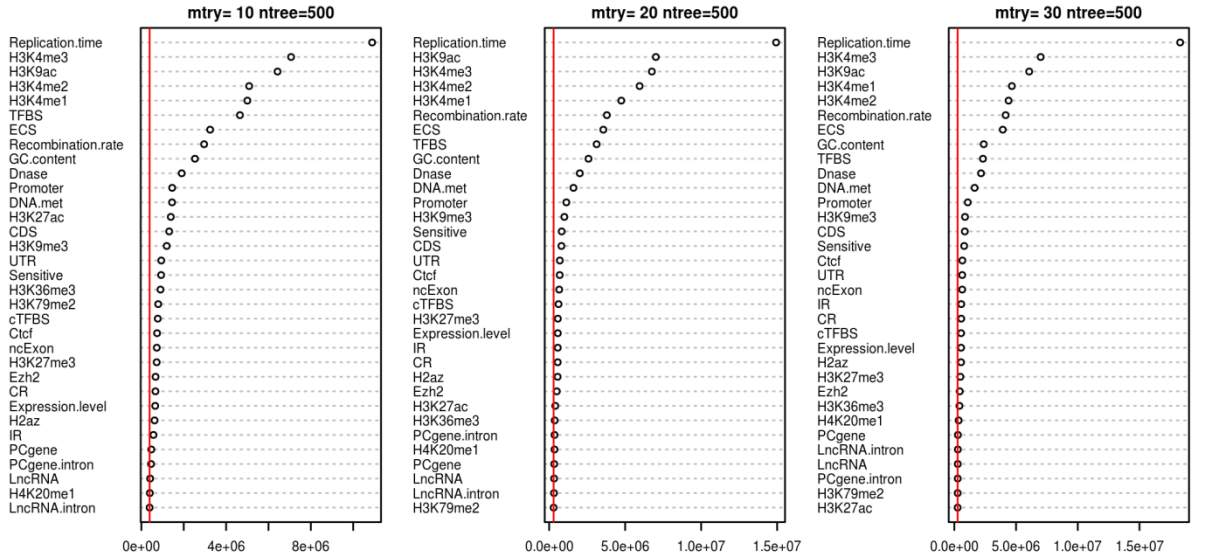

Figure S19. variable importance (lncNodePurity) sensitivity to ntree and mtry (red line: the absolute value of minimum importance among all features in the SOM model of liver cancer)

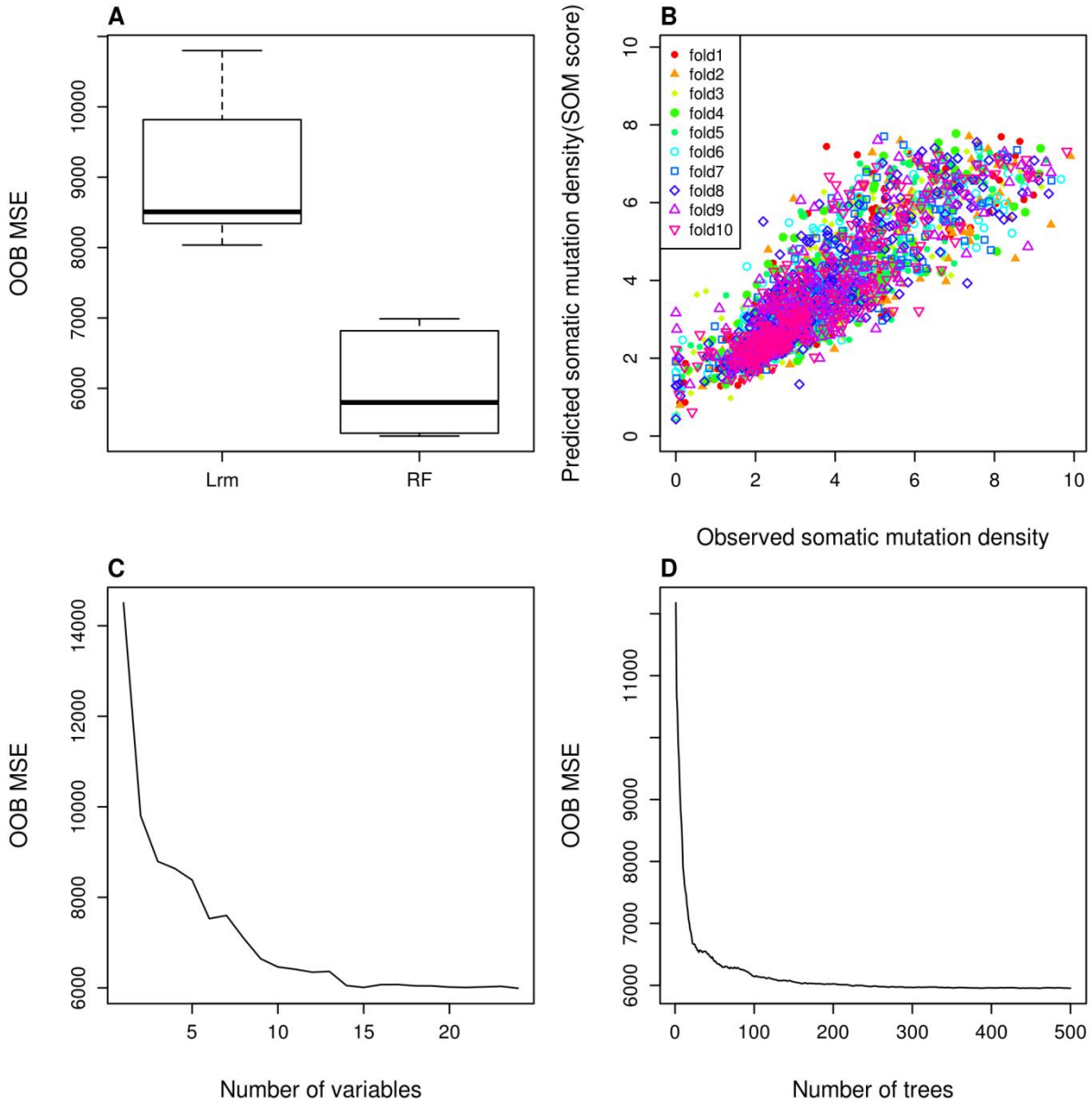

Figure S20. Validation of the liver cancer SOM model. A.MSE for linear regression model (Lrm) and Random forest (RF) with 10-fold cross validation (SOM model); B. observed and predicted somatic mutation density divided by 88 patients with 10-fold cross validation; C. the number of variables remained in the RF model minimizes the OOB error; D. the number of trees in the RF model minimizing the OOB error.

6.4 1-Publication in Cancer Letters

Cancer Letters 369 (2015) 307–315

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Mini-review

Mining the coding and non-coding genome for cancer drivers

Jia Li [a], Damien Drubay [b,c], Stefan Michiels [b,c], Daniel Gautheret [a,*]

[a] Institute for Integrative Biology of the Cell (I2BC), CNRS, CEA, Université Paris-Sud, Université Paris-Saclay, 91198 Gif sur Yvette, France
[b] Service de Biostatistique et d'Epidemiologie, Gustave Roussy, Villejuif, France
[c] INSERM U1018, CESP, Université Paris-Sud, Université Paris-Saclay, Villejuif, France

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form 24 September 2015
Accepted 24 September 2015

Keywords:
Cancer drivers
Non-coding drivers
Somatic mutation scoring
Bioinformatics

ABSTRACT

Progress in next-generation sequencing provides unprecedented opportunities to fully characterize the spectrum of somatic mutations of cancer genomes. Given the large number of somatic mutations identified by such technologies, the prioritization of cancer-driving events is a consistent bottleneck. Most bioinformatics tools concentrate on driver mutations in the coding fraction of the genome, those causing changes in protein products. As more non-coding pathogenic variants are identified and characterized, the development of computational approaches to effectively prioritize cancer-driving variants within the non-coding fraction of human genome is becoming critical. After a short summary of methods for coding variant prioritization, we here review the highly diverse non-coding elements that may act as cancer drivers and describe recent methods that attempt to evaluate the deleteriousness of sequence variation in these elements. With such tools, the prioritization and identification of cancer-implicated regulatory elements and non-coding RNAs is becoming a reality.

Introduction

Cancer is caused by the accumulation of genetic alterations and consequent disruption of cell functions. Over the past decade, the introduction of fast and relatively inexpensive sequencing methods has provided unprecedented opportunity to characterize cancer genomic landscapes. A variety of bioinformatics tools are now available to discover genetic variations from high throughput sequencing of tumor DNA, such as GATK [1], CRISP [2], LoFreq [3], VarScan 2 [4], and SNVer [5], which have been recently evaluated [6,7]. Depending on cancer type, tumors harbor hundreds to tens of thousands of somatic mutations, most of which are located in the non-coding portion of the genome [8]. A critical challenge in this context is to distinguish "driver" mutations and cancer genes that actively contribute to tumor growth or metastasis from "passenger" mutations that are mere results of the cancerous process. A number of reviews provide guidelines for the discovery of cancer-causing variants [9,10]. The most common strategy is first to prioritize non-synonymous variants in protein-coding regions and then seek recurrently mutated genes in a cohort of cancer patients [11–15]. Diverse computational methods have been explored to prioritize non-synonymous variants with respect to their disease-causing potential. Most are based on the assumption that coding mutations impacting functionally important residues, as inferred from evolutionary conservation and protein domain analysis, are more likely damaging [16]. Other software, used in conjunction with these scoring systems, perform recurrence search in patient cohorts. Currently, 547 cancer genes are described the COSMIC catalogue of somatic mutations in cancer (version 71) [17].

The immense majority of the human genome (98%) is non-coding, and consequently most somatic mutations/alterations observed in tumors occur in this non-coding fraction. Because non-coding mutations are more difficult to interpret, these regions have been mostly discounted from the wider search for driver mutations. However, mutations in non-coding regions can have a profound impact on cell fate. Indeed, functional regions in the non-coding genome include mRNA splice sites, UTR regulation elements, promoters, transcription factor binding sites, enhancers and a wide variety of non-coding RNA (ncRNA) genes. Among ncRNA genes, one particular class is now receiving focused attention due to its vast extent: long non-coding RNA (lncRNA). According to the latest estimate [18], over 58,000 lncRNA genes are expressed in the human genome, which makes this class the biggest contributor to the "black matter" transcriptome.

There is ample evidence for disease-related mutations in the non-coding genome. A large fraction of disease or trait-relevant single nucleotide polymorphisms (SNPs) detected by Genome-wide Association Studies (GWAS) [19] is located in the non-coding genome, preferentially within enhancers, exons and mRNA promoters [20]. Inherited disease-causing variants are strongly enriched in non-coding regions under strong purifying selection, which comprise binding sites of transcription factors (TFs) and critical motifs from TF Families [21]. Further studies have shown that altered ncRNA

* Corresponding author. Tel.: +33169154632; fax: +33169157296.
E-mail address: daniel.gautheret@u-psud.fr (D. Gautheret).

http://dx.doi.org/10.1016/j.canlet.2015.09.015
0304-3835/

6.5 2-Publication in PLoS Computational Biology

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A Dual Model for Prioritizing Cancer Mutations in the Non-coding Genome Based on Germline and Somatic Events

Jia Li[1], Marie-Anne Poursat[2], Damien Drubay[3,4], Arnaud Motz[1], Zohra Saci[5], Antonin Morillon[5], Stefan Michiels[3,4], Daniel Gautheret[1]*

1 Institute for Integrative Biology of the Cell, Université Paris-Sud, Paris, France, **2** Laboratoire de Mathématique, Université Paris-Sud, Paris, France, **3** Service de Biostatistique et d'Epidémiologie, Gustave Roussy, Villejuif, France, **4** INSERM U1018, CESP, Université Paris-Sud, Villejuif, France, **5** RNA, epigenetics and genome fluidity, Institut Curie, PSL Research University, CNRS UMR3244, Université Pierre et Marie Curie, Paris, France

* daniel.gautheret@u-psud.fr

OPEN ACCESS

Citation: Li J, Poursat M-A, Drubay D, Motz A, Saci Z, Morillon A, et al. (2015) A Dual Model for Prioritizing Cancer Mutations in the Non-coding Genome Based on Germline and Somatic Events. PLoS Comput Biol 11(11): e1004583. doi:10.1371/journal.pcbi.1004583

Editor: Rachel Karchin, Johns Hopkins University, UNITED STATES

Received: June 11, 2015

Accepted: October 4, 2015

Published: November 20, 2015

Data Availability Statement: Other supplementary files are made publicly available through our repository: http://biodev.cea.fr/98drivers.

Funding: This project was funded in part by "Plan Cancer – Systems Biology" grant #bio2014-04 to DG, SM and Amor. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

We address here the issue of prioritizing non-coding mutations in the tumoral genome. To this aim, we created two independent computational models. The first (germline) model estimates purifying selection based on population SNP data. The second (somatic) model estimates tumor mutation density based on whole genome tumor sequencing. We show that each model reflects a different set of constraints acting either on the normal or tumor genome, and we identify the specific genome features that most contribute to these constraints. Importantly, we show that the somatic mutation model carries independent functional information that can be used to narrow down the non-coding regions that may be relevant to cancer progression. On this basis, we identify positions in non-coding RNAs and the non-coding parts of mRNAs that are both under purifying selection in the germline and protected from mutation in tumors, thus introducing a new strategy for future detection of cancer driver elements in the expressed non-coding genome.

Author Summary

Cancer cells undergo a mutation/selection process that resembles that of any living cell. Most mutations in cancer cell DNA occur in the so-called "non-coding" regions that represent 98.5% of the genome length. Pinning down which of these mutations contribute to the fitness of cancer cells would be important for identifying new "cancer drivers", which may in turn lead to future treatments. Unfortunately, predicting the impact of a non-coding DNA alteration remains extremely difficult. In this study, we analyze millions of non-coding cancer mutations and show cancer-specific mutational patterns can be used to predict non-coding regions that are preserved from mutations and may thus be important for cancer cell survival. Combining this information with population data, we propose a

Reference

Adzhubei, I. a, Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. doi:10.1038/nmeth0410-248

Alexander, N.R., Tran, N.L., Rekapally, H., Summers, C.E., Glackin, C., Heimark, R.L., 2006. N-cadherin gene expression in prostate carcinoma is modulated by integrin-dependent nuclear translocation of Twist1. Cancer Res. 66, 3365–3369. doi:10.1158/0008-5472.CAN-05-3401

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S. a J.R., Behjati, S., Biankin, A. V, Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjörd, J.E., Foekens, J. a, Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jäger, N., Jones, D.T.W., Jones, D., Knappskog, S., Kool, M., Lakhani, S.R., López-Otín, C., Martin, S., Munshi, N.C., Nakamura, H., Northcott, P. a, Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V, Puente, X.S., Raine, K., Ramakrishna, M., Richardson, A.L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T.N., Span, P.N., Teague, J.W., Totoki, Y., Tutt, A.N.J., Valdés-Mas, R., van Buuren, M.M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L.R., Zucman-Rossi, J., Futreal, P.A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S.M., Siebert, R., Campo, E., Shibata, T., Pfister, S.M., Campbell, P.J., Stratton, M.R., 2013. Signatures of mutational processes in human cancer. Nature 500, 415–21. doi:10.1038/nature12477

Altshuler, D.M., Gibbs, R. a, Peltonen, L., Altshuler, D.M., Gibbs, R. a, Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P.E., Altshuler, D.M., Gibbs, R. a, de Bakker, P.I.W., Deloukas, P., Gabriel, S.B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L.R., Ren, Y., Wheeler, D., Gibbs, R. a, Muzny, D.M., Barnes, C., Darvishi, K., Hurles, M., Korn, J.M., Kristiansson, K., Lee, C., McCarrol, S. a, Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S.B., Pollack, S., Price, A.L., Soranzo, N., Bonnen, P.E., Gibbs, R. a, Gonzaga-Jauregui, C., Keinan, A., Price, A.L., Yu, F., Anttila, V., Brodeur, W., Daly, M.J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S.F., Zhang, Q., Ghori, M.J.R., McGinnis, R., McLaren, W., Pollack, S., Price, A.L., Schaffner, S.F., Takeuchi, F., Grossman, S.R., Shlyakhter, I., Hostetter, E.B., Sabeti, P.C., Adebamowo, C. a, Foster, M.W., Gordon, D.R., Licinio, J., Manca, M.C., Marshall, P. a, Matsuda, I., Ngare, D., Wang, V.O., Reddy, D., Rotimi, C.N., Royal, C.D., Sharp, R.R., Zeng, C., Brooks, L.D., McEwen, J.E., 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58. doi:10.1038/nature09298

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, a M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D. a, Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M., Sandelin, A., 2014. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–61. doi:10.1038/nature12787

Askarian-Amiri, M.E., Seyfoddin, V., Smart, C.E., Wang, J., Kim, J.E., Hansji, H., Baguley, B.C., Finlay, G.J., Leung, E.Y., 2014. Emerging role of long non-coding RNA SOX2OT in SOX2 regulation in breast cancer. PLoS One 9, 1–10. doi:10.1371/journal.pone.0102140

Bansal, V., 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 26, 318–324. doi:10.1093/bioinformatics/btq214

Bao, L., Zhou, M., Cui, Y., 2005. nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 33, 480–482. doi:10.1093/nar/gki372

Barsyte-Lovejoy, D., 2006. The c-Myc Oncogene Directly Induces the H19 Noncoding RNA by Allele-Specific Binding to Potentiate Tumorigenesis. Cancer Res. 66, 5330–5337. doi:10.1158/0008-5472.CAN-06-0037

Beck, T., Hastings, R.K., Gollapudi, S., Free, R.C., Brookes, A.J., 2014. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur. J. Hum. Genet. 22, 949–52. doi:10.1038/ejhg.2013.274

Bellucci, M., Agostini, F., Masin, M., Tartaglia, G.G., 2011. Predicting protein associations with long noncoding RNAs. Nat. Methods 8, 444–445. doi:10.1038/nmeth.1611

Beltran, M., Puig, I., Peña, C., Garcá, J.M., Álvarez, A.B., Peña, R., Bonilla, F., Herreros, A.G. De, 2008. A natural antisense transcript regulates Zeb2 / Sip1 gene expression during Snail1-induced epithelial – mesenchymal transition. Genes Dev. 22, 756–769. doi:10.1101/gad.455708.in

Benetatos, L., Vartholomatos, G., Hatzimichael, E., 2011. MEG3 imprinted gene contribution in tumorigenesis. Int. J. Cancer 129, 773–779. doi:10.1002/ijc.26052

Berger, M.F., Hodis, E., Heffernan, T.P., Deribe, Y.L., Lawrence, M.S., Protopopov, A., Ivanova, E., Watson, I.R., Nickerson, E., Ghosh, P., Zhang, H., Zeid, R., Ren, X., Cibulskis, K., Sivachenko, A.Y., Wagle, N., Sucker, A., Sougnez, C., Onofrio, R., Ambrogio, L., Auclair, D., Fennell, T., Carter, S.L., Drier, Y., Stojanov, P., Singer, M. a., Voet, D., Jing, R., Saksena, G., Barretina, J., Ramos, A.H., Pugh, T.J., Stransky, N., Parkin, M., Winckler, W., Mahan, S., Ardlie, K., Baldwin, J., Wargo, J., Schadendorf, D., Meyerson, M., Gabriel, S.B., Golub, T.R., Wagner, S.N., Lander, E.S., Getz, G., Chin, L., Garraway, L. a., 2012. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature 485, 502–506. doi:10.1038/nature11071

Bernstein, B.E., Stamatoyannopoulos, J. a, Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. a, Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J. a, 2010. The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. 28, 1045–1048. doi:10.1038/nbt1010-1045

Berteaux, N., Lottin, S., Monté, D., Pinte, S., Quatannens, B., Coll, J., Hondermarck, H., Curgy, J.J., Dugimont, T., Adriaenssens, E., 2005. H19 mRNA-like noncoding RNA

promotes breast cancer cell proliferation through positive control by E2F1. J. Biol. Chem. 280, 29625–29636. doi:10.1074/jbc.M504033200

Berx, G., van Roy, F., 2009. Involvement of Members of the Cadherin Superfamily in Cancer. Cold Spring Harb. Perspect. Biol. 1, a003129–a003129. doi:10.1101/cshperspect.a003129

Bida, O., Gidoni, M., Ideses, D., Efroni, S., Ginsberg, D., 2015. A novel mitosis-associated lncRNA , MA-linc1 , is required for cell cycle progression and sensitizes cancer cells to Paclitaxel.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370. doi:10.1093/nar/gkg095

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M. a., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., Cherry, J.M., Snyder, M., 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22, 1790–1797. doi:10.1101/gr.137323.112

Braconi, C., Valeri, N., Kogure, T., Gasparini, P., Huang, N., Nuovo, G.J., Terracciano, L., Croce, C.M., Patel, T., 2010. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. Proc. Natl. Acad. Sci. 108, 786–791. doi:10.1073/pnas.1011098108

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324

Bromberg, Y., Rost, B., 2007. SNAP: Predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 35, 3823–3835. doi:10.1093/nar/gkm238

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L., Casadio, R., 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum. Mutat. 30, 1237–1244. doi:10.1002/humu.21047

Calin, G. a., Liu, C., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E.J., Wojcik, S.E., Shimizu, M., Tili, E., Rossi, S., Taccioli, C., Pichiorri, F., Liu, X., Zupo, S., Herlea, V., Gramantieri, L., Lanza, G., Alder, H., Rassenti, L., Volinia, S., Schmittgen, T.D., Kipps, T.J., Negrini, M., Croce, C.M., 2007. Ultraconserved Regions Encoding ncRNAs Are Altered in Human Leukemias and Carcinomas. Cancer Cell 12, 215–229. doi:10.1016/j.ccr.2007.07.027

Capriotti, E., Calabrese, R., Casadio, R., 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22, 2729–2734. doi:10.1093/bioinformatics/btl423

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Karchin, R., 2009. Cancer-specific high-throughput annotation of somatic mutations:

Computational prediction of driver missense mutations. Cancer Res. 69, 6660–6667. doi:10.1158/0008-5472.CAN-09-1133

Cavallaro, U., Christofori, G., 2004. Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. Nat. Rev. Cancer 4, 118–132. doi:10.1038/nrc1276

Chaluvally-Raghavan, P., Zhang, F., Pradeep, S., Hamilton, M.P., Zhao, X., Rupaimoole, R., Moss, T., Lu, Y., Yu, S., Pecot, C.V., Aure, M.R., Peuget, S., Rodriguez-Aguayo, C., Han, H.-D., Zhang, D., Venkatanarayan, A., Krohn, M., Kristensen, V.N., Gagea, M., Ram, P., Liu, W., Lopez-Berestein, G., Lorenzi, P.L., Børresen-Dale, A.-L., Chin, K., Gray, J., Dusetti, N.J., McGuire, S.E., Flores, E.R., Sood, A.K., Mills, G.B., 2014. Copy Number Gain of hsa-miR-569 at 3q26.2 Leads to Loss of TP53INP1 and Aggressiveness of Epithelial Cancers. Cancer Cell 26, 863–879. doi:10.1016/j.ccell.2014.10.010

Chapman, M. a, Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.-P., Ahmann, G.J., Adli, M., Anderson, K.C., Ardlie, K.G., Auclair, D., Baker, A., Bergsagel, P.L., Bernstein, B.E., Drier, Y., Fonseca, R., Gabriel, S.B., Hofmeister, C.C., Jagannath, S., Jakubowiak, A.J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L.M., Onofrio, R., Pugh, T.J., Rajkumar, S.V., Ramos, A.H., Siegel, D.S., Sivachenko, A., Stewart, a K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W.C., Garraway, L. a, Meyerson, M., Lander, E.S., Getz, G., Golub, T.R., 2011. Initial genome sequencing and analysis of multiple myeloma. Nature 471, 467–472. doi:10.1038/nature09837

Chen, C.-L., Tseng, Y.-W., Wu, J.-C., Chen, G.-Y., Lin, K.-C., Hwang, S.-M., Hu, Y.-C., 2015. Suppression of hepatocellular carcinoma by baculovirus-mediated expression of long non-coding RNA PTENP1 and MicroRNA regulation. Biomaterials 44, 71–81. doi:10.1016/j.biomaterials.2014.12.023

Chen, K., Rajewsky, N., 2006. Natural selection on human microRNA binding sites inferred from SNP data. Nat. Genet. 38, 1452–1456. doi:10.1038/ng1910

Chen, W., Huang, M., Kong, R., Xu, T., Zhang, E., Xia, R., Sun, M., De, W., Shu, Y., 2015. Antisense Long Noncoding RNA HIF1A-AS2 Is Upregulated in Gastric Cancer and Associated with Poor Prognosis. Dig. Dis. Sci. doi:10.1007/s10620-015-3524-0

Chen, Y., Cunningham, F., Rios, D., McLaren, W.M., Smith, J., Pritchard, B., Spudich, G.M., Brent, S., Kulesha, E., Marin-Garcia, P., Smedley, D., Birney, E., Flicek, P., 2010. Ensembl variation resources. BMC Genomics 11, 293. doi:10.1186/1471-2164-11-293

Cheng, Y., Jutooru, I., Chadalapaka, G., Corton, J.C., Safe, S., 2015. The long non-coding RNA HOTTIP enhances pancreatic cancer cell proliferation, survival and migration. Oncotarget 6, 10840–52.

Chipuk, J.E., Kuwana, T., Bouchier-Hayes, L., Droin, N.M., Newmeyer, D.D., Schuler, M., Green, D.R., 2004. Direct activation of Bax by p53 mediates mitochondrial membrane permeabilization and apoptosis. Science 303, 1010–1014. doi:10.1126/science.1092734

Chung, S., Nakagawa, H., Uemura, M., Piao, L., Ashikawa, K., Hosono, N., Takata, R., Akamatsu, S., Kawaguchi, T., Morizono, T., Tsunoda, T., Daigo, Y., Matsuda, K., Kamatani, N., Nakamura, Y., Kubo, M., 2011. Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. Cancer Sci. 102, 245–252. doi:10.1111/j.1349-7006.2010.01737.x

Clark, M.B., Mattick, J.S., 2011. Long noncoding RNAs in cell biology. Semin. Cell Dev. Biol. 22, 366–376. doi:10.1016/j.semcdb.2011.01.001

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M., Sherry, S., Flicek, P., 2012. The 1000 Genomes Project: data management and community access. Nat. Methods 9, 459–462. doi:10.1038/nmeth.1974

Clemson, C.M., Hutchinson, J.N., Sara, S. a., Ensminger, A.W., Fox, A.H., Chess, A., Lawrence, J.B., 2009. An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles. Mol. Cell 33, 717–726. doi:10.1016/j.molcel.2009.01.026

Clifford, R.J., Edmonson, M.N., Nguyen, C., Buetow, K.H., 2004. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics 20, 1006–1014. doi:10.1093/bioinformatics/bth029

Coccia, E.M., Cicala, C., Charlesworth, a, Ciccarelli, C., Rossi, G.B., Philipson, L., Sorrentino, V., 1992. Regulation and expression of a growth arrest-specific gene (gas5) during growth, differentiation, and development. Mol. Cell. Biol. 12, 3514–21.

Corley, M., Solem, a., Qu, K., Chang, H.Y., Laederach, a., 2015. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. Nucleic Acids Res. 43, 1859–1868. doi:10.1093/nar/gkv010

Cunnington, M.S., Santibanez Koref, M., Mayosi, B.M., Burn, J., Keavney, B., 2010. Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. PLoS Genet. 6, e1000899. doi:10.1371/journal.pgen.1000899

D'Antonio, M., Ciccarelli, F.D., 2013. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol. 14, R52. doi:10.1186/gb-2013-14-5-r52

Dawson, M. a., Kouzarides, T., 2012. Cancer epigenetics: From mechanism to therapy. Cell 150, 12–27. doi:10.1016/j.cell.2012.06.013

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., Wilson, R.K., Ding, L., 2012. MuSiC: Identifying mutational significance in cancer genomes. Genome Res. 22, 1589–1598. doi:10.1101/gr.134635.111

Degner, J.F., Pai, A. a., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., Stephens, M., Gilad, Y., Pritchard,

J.K., 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390–394. doi:10.1038/nature10808

Delgado André, N., De Lucca, F.L., 2008. Non-coding transcript in T cells (NTT): Antisense transcript activates PKR and NF-κB in human lymphocytes. Blood Cells, Mol. Dis. 40, 227–232. doi:10.1016/j.bcmd.2007.08.005

DeOcesano-Pereira, C., Amaral, M.S., Parreira, K.S., Ayupe, a. C., Jacysyn, J.F., Amarante-Mendes, G.P., Reis, E.M., Verjovski-Almeida, S., 2014. Long non-coding RNA INXS is a critical mediator of BCL-XS induced apoptosis. Nucleic Acids Res. 42, 8343–8355. doi:10.1093/nar/gku561

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A. a, del Angel, G., Rivas, M. a, Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498. doi:10.1038/ng.806

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C. a, Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., Guigó, R., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22, 1775–89. doi:10.1101/gr.132159.111

Ding, L., Getz, G., Wheeler, D. a, Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., Fulton, L., Fulton, R.S., Zhang, Q., Wendl, M.C., Lawrence, M.S., Larson, D.E., Chen, K., Dooling, D.J., Sabo, A., Hawes, A.C., Shen, H., Jhangiani, S.N., Lewis, L.R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S.E., Clerc, K., Metcalf, G. a, Ng, B., Milosavljevic, A., Gonzalez-Garay, M.L., Osborne, J.R., Meyer, R., Shi, X., Tang, Y., Koboldt, D.C., Lin, L., Abbott, R., Miner, T.L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B.H., Kraja, A., Crosby, S.D., Sawyer, C.S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L.R., Dutt, A., Fennell, T., Hanna, M., Johnson, B.E., Onofrio, R.C., Thomas, R.K., Tonon, G., Weir, B. a, Zhao, X., Ziaugra, L., Zody, M.C., Giordano, T., Orringer, M.B., Roth, J. a, Spitz, M.R., Wistuba, I.I., Ozenberger, B., Good, P.J., Chang, A.C., Beer, D.G., Watson, M. a, Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W.D., Pao, W., Province, M. a, Weinstock, G.M., Varmus, H.E., Gabriel, S.B., Lander, E.S., Gibbs, R. a, Meyerson, M., Wilson, R.K., 2008. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455, 1069–1075. doi:10.1038/nature07423

Dobin, A., Davis, C. a, Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. doi:10.1093/bioinformatics/bts635

Du, Z., Fei, T., Verhaak, R.G.W., Su, Z., Zhang, Y., Brown, M., Chen, Y., Liu, X.S., 2013. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat. Struct. Mol. Biol. 20, 908–913. doi:10.1038/nsmb.2591

Facts, C., 2015. American Cancer Society: Cancer Facts and Figures 2015. doi:10.3322/caac.21254

Fang, Z., Wu, L., Wang, L., Yang, Y., Meng, Y., Yang, H., 2014. Increased expression of the long non-coding RNA UCA1 in tongue squamous cell carcinomas: a possible correlation with cancer metastasis. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 117, 89–95. doi:10.1016/j.oooo.2013.09.007

Feitelson, M. a., Arzumanyan, A., Kulathinal, R.J., Blain, S.W., Holcombe, R.F., Mahajna, J., Marino, M., Martinez-Chantar, M.L., Nawroth, R., Sanchez-Garcia, I., Sharma, D., Saxena, N.K., Singh, N., Vlachostergios, P.J., Guo, S., Honoki, K., Fujii, H., Georgakilas, A.G., Amedei, A., Niccolai, E., Amin, A., Ashraf, S.S., Boosani, C.S., Guha, G., Ciriolo, M.R., Aquilano, K., Chen, S., Mohammed, S.I., Azmi, A.S., Bhakta, D., Halicka, D., Nowsheen, S., 2015. Sustained proliferation in cancer: Mechanisms and novel therapeutic targets. Semin. Cancer Biol. 1–30. doi:10.1016/j.semcancer.2015.02.006

Fellig, Y., Ariel, I., Ohana, P., Schachter, P., Sinelnikov, I., Birman, T., Ayesh, S., Schneider, T., de Groot, N., Czerniak, a, Hochberg, a, 2005. H19 expression in hepatic metastases from a range of human carcinomas. J. Clin. Pathol. 58, 1064–1068. doi:10.1136/jcp.2004.023648

Feng, J., Bi, C., Clark, B.S., Mady, R., Shah, P., Kohtz, J.D., 2006. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev. 20, 1470–1484. doi:10.1101/gad.1416106

Flockhart, R.J., Webster, D.E., Qu, K., Mascarenhas, N., Kovalski, J., Kretz, M., Khavari, P. a, 2012. BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. Genome Res. 22, 1006–14. doi:10.1101/gr.140061.112

Forbes, S. a., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J.W., Campbell, P.J., Stratton, M.R., Futreal, P.A., 2011a. COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 39, 945–950. doi:10.1093/nar/gkq929

Forbes, S. a., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J.W., Campbell, P.J., Stratton, M.R., Futreal, P.A., 2011b. COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 39, 945–950. doi:10.1093/nar/gkq929

French, B.N., Et Al, 2013. Functional Variants at the 11q13 Breast Cancer Risk Loci Regulate Cyclin D1 Expression through Long-Range Enhancers 92, 78540526. doi:10.1016/j.ajhg.2013.01.002.

Frousios, K., Iliopoulos, C.S., Schlitt, T., Simpson, M. a., 2013. Predicting the functional consequences of non-synonymous DNA sequence variants - evaluation of bioinformatics tools and development of a consensus strategy. Genomics 102, 223–228. doi:10.1016/j.ygeno.2013.06.005

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., Gerstein, M., 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 15, 1–15. doi:10.1186/s13059-014-0480-5

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R., 2004. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183. doi:10.1038/nrc1299

Gallia, G.L., Johnson, E.M., Khalili, K., 2000. Puralpha: a multifunctional single-stranded DNA- and RNA-binding protein. Nucleic Acids Res. 28, 3197–3205.

Garding, A., Bhattacharya, N., Haebe, S., Müller, F., Weichenhan, D., Idler, I., Ickstadt, K., Stilgenbauer, S., Mertens, D., 2013. TCL1A and ATM are co-expressed in chronic lymphocytic leukemia cells without deletion of 11q. Haematologica 98, 269–73. doi:10.3324/haematol.2012.070623

Geng, Y.J., Xie, S.L., Li, Q., Ma, J., Wang, G.Y., 2011. Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. J. Int. Med. Res. 39, 2119–28. doi:10.1177/147323001103900608

Genuer, R., Poggi, J., Tuleau-malot, C., Genuer, R., Poggi, J., Variable, C.T., Genuer, R., Poggi, J., Tuleau-malot, C., 2012. Variable selection using Random Forests. PATTERN RECOGN LETT 31, 2225–2236.

Gezer, U., Ph, D., Tiryakioglu, D., Sc, M., Bilgin, E., Sc, M., Dalay, N., Ph, D., 2015. Androgen Stimulation of PCA3 and miR-141 and Their Release from Prostate Cancer Cells 16, 488–493.

Giardine, B., Riemer, C., Hefferon, T., Thomas, D., Hsu, F., Zielenski, J., Sang, Y., Elnitski, L., Cutting, G., Trumbower, H., Kern, A., Kuhn, R., Patrinos, G.P., Hughes, J., Higgs, D., Chui, D., Scriver, C., Phommarinh, M., Patnaik, S.K., Blumenfeld, O., Gottlieb, B., Vihinen, M., Väliaho, J., Kent, J., Miller, W., Hardison, R.C., 2007. PhenCode: Connecting ENCODE data with mutations and phenotype. Hum. Mutat. 28, 554–562. doi:10.1002/humu.20484

Goldar, S., Khaniani, M.S., Derakhshan, S.M., Baradaran, B., 2015. Molecular mechanisms of apoptosis and roles in cancer development and treatment. Asian Pac. J. Cancer Prev. 16, 2129–2144.

Gonzalez-Perez, A., Lopez-Bigas, N., 2012. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 40, 1–10. doi:10.1093/nar/gks743

González-Pérez, A., López-Bigas, N., 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. 88, 440–449. doi:10.1016/j.ajhg.2011.03.004

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., Lopez-Bigas, N., 2013. IntOGen-mutations identifies cancer drivers across tumor types. Nat. Methods 10, 1081–2. doi:10.1038/nmeth.2642

Gopalakrishnan, C., Kamaraj, B., Purohit, R., 2014. Mutations in microRNA Binding Sites of CEP Genes Involved in Cancer. Cell Biochem. Biophys. 70, 1–10. doi:10.1007/s12013-014-0153-8

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864. doi:10.1126/science.185.4154.862

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E.E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D.P., Louis, D.N., Goldstraw, P., Nicholson, A.G., Brasseur, F., Looijenga, L., Weber, B.L., Chiew, Y.-E., DeFazio, A., Greaves, M.F., Green, A.R., Campbell, P., Birney, E., Easton, D.F., Chenevix-Trench, G., Tan, M.-H., Khoo, S.K., Teh, B.T., Yuen, S.T., Leung, S.Y., Wooster, R., Futreal, P.A., Stratton, M.R., 2007. Patterns of somatic mutation in human cancer genomes. Nature 446, 153–158. doi:10.1038/nature05610

Gross, D.S., Garrard, W.T., 1988. Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. 57, 159–197. doi:10.1146/annurev.biochem.57.1.159

Gui, Y., Guo, G., Huang, Y., Hu, X., Tang, A., Gao, S., Wu, R., Chen, C., Li, X., Zhou, L., He, M., Li, Z., Sun, X., Jia, W., Chen, J., Yang, S., Zhou, F., Zhao, X., Wan, S., Ye, R., Liang, C., Liu, Z., Huang, P., Liu, C., Jiang, H., Wang, Y., Zheng, H., Sun, L., Liu, X., Jiang, Z., Feng, D., Chen, J., Wu, S., Zou, J., Zhang, Z., Yang, R., Zhao, J., Xu, C., Yin, W., Guan, Z., Ye, J., Zhang, H., Li, J., Kristiansen, K., Nickerson, M.L., Theodorescu, D., Li, Y., Zhang, X., Li, S., Wang, J., Yang, H., Wang, J., Cai, Z., 2011. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nat. Genet. 43, 875–878. doi:10.1038/ng.907

Gumireddy, K., Li, A., Yan, J., Setoyama, T., Johannes, G.J., Orom, U. a, Tchou, J., Liu, Q., Zhang, L., Speicher, D.W., Calin, G. a, Huang, Q., 2013. Identification of a long non-coding RNA-associated RNP complex regulating metastasis at the translational step. EMBO J. 32, 2672–84. doi:10.1038/emboj.2013.188

Guo, F., Li, Y., Liu, Y., Wang, J., Li, Y., Li, G., 2010. Inhibition of metastasis-associated lung adenocarcinoma transcript 1 in CaSki human cervical cancer cells suppresses cell proliferation and invasion. Acta Biochim. Biophys. Sin. (Shanghai). 42, 224–229. doi:10.1093/abbs/gmq008.Inhibition

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., Luo, H., Zhao, G., Bu, D., Jiao, F., Shao, Q., Chen, R., Zhao, Y., 2013. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic Acids Res. 41, e35–e35. doi:10.1093/nar/gks967

Gupta, R. a, Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R.B., van de Vijver, M.J., Sukumar, S., Chang, H.Y., 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076. doi:10.1038/nature08975

Gutschner, T., Baas, M., Diederichs, S., 2011. Noncoding RNA gene silencing through genomic integration of RNA destabilizing elements using zinc finger nucleases. Genome Res. 21, 1944–54. doi:10.1101/gr.122358.111

Gutschner, T., Diederichs, S., 2012. The hallmarks of cancer. RNA Biol. 9, 703–719. doi:10.4161/rna.20481

Gutschner, T., Hämmerle, M., Eißmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Groß, M., Zörnig, M., MacLeod, a. R., Spector, D.L., Diederichs, S., 2013. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. 73, 1180–1189. doi:10.1158/0008-5472.CAN-12-2850

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., Lander, E.S., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227. doi:10.1038/nature07672

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E., Lander, E.S., 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295–300. doi:10.1038/nature10398

Haerty, W., Ponting, C.P., 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome Biol 14, R49. doi:10.1186/gb-2013-14-5-r49

Han, L., Zhang, E., Yin, D., Kong, R., Xu, T., Chen, W., Xia, R., Shu, Y., De, W., 2015. Low expression of long noncoding RNA PANDAR predicts a poor prognosis of non-small cell lung cancer and affects cell apoptosis by regulating Bcl-2. Cell Death Dis. 6, e1665. doi:10.1038/cddis.2015.30

Hanahan, D., Weinberg, R. a, 2011. Hallmarks of cancer: the next generation. Cell 144, 646–74. doi:10.1016/j.cell.2011.02.013

Hao, Y., Wu, W., Shi, F., Dalmolin, R.J., Yan, M., Tian, F., Chen, X., Chen, G., Cao, W., 2015. Prediction of long noncoding RNA functions with co-expression network in esophageal squamous cell carcinoma. BMC Cancer 15, 168. doi:10.1186/s12885-015-1179-z

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., Van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., Hubbard, T.J., 2012. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 22, 1760–1774. doi:10.1101/gr.135350.111

He, C., Zhou, F., Zuo, Z., Cheng, H., Zhou, R., 2009. A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis. PLoS One 4. doi:10.1371/journal.pone.0004732

He, Y., Carrillo, J. a., Luo, J., Ding, Y., Tian, F., Davidson, I., Song, J., 2014. Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells. Front. Genet. 5, 1–9. doi:10.3389/fgene.2014.00308

Hodgkinson, A., Chen, Y., Eyre-Walker, A., 2012. The large-scale distribution of somatic mutations in cancer genomes. Hum. Mutat. 33, 136–143. doi:10.1002/humu.21616

Hodis, E., Watson, I.R., Kryukov, G. V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., Auclair, D., Li, L., Place, C., Dicara, D., Ramos, A.H., Lawrence, M.S., Cibulskis, K., Sivachenko, A., Voet, D., Saksena, G., Stransky, N., Onofrio, R.C., Winckler, W., Ardlie, K., Wagle, N., Wargo, J., Chong, K., Morton, D.L., Stemke-Hale, K., Chen, G., Noble, M., Meyerson, M., Ladbury, J.E., Davies, M. a., Gershenwald, J.E., Wagner, S.N., Hoon, D.S.B., Schadendorf, D., Lander, E.S., Gabriel, S.B., Getz, G., Garraway, L. a., Chin, L., 2012. A landscape of driver mutations in melanoma. Cell 150, 251–263. doi:10.1016/j.cell.2012.06.024

Hollander, M.C., Alamo, I., Fornace Jr., a J., 1996. A novel DNA damage-inducible transcript, gadd7, inhibits cell growth, but lacks a protein product. Nucleic Acids Res 24, 1589–1593. doi:6g0105 [pii]

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., Schadendorf, D., Kumar, R., 2013. TERT promoter mutations in familial and sporadic melanoma. Science 339, 959–61. doi:10.1126/science.1230062

Hou, Z., Zhao, W., Zhou, J., Shen, L., Zhan, P., Xu, C., Chang, C., Bi, H., Zou, J., Yao, X., Huang, R., Yu, L., Yan, J., 2014. A long noncoding RNA Sox2ot regulates lung cancer cell proliferation and is a prognostic indicator of poor survival. Int. J. Biochem. Cell Biol. 53, 380–388. doi:10.1016/j.biocel.2014.06.004

Hs, N., Tr, H., Morris, Q., Barash, Y., Ar, K., Jojic, N., Sw, S., 2015. RNA splicing . The human splicing code reveals new insights into the genetic determinants of disease . 347, 1254806. doi:10.1126/science.1254806.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G. V, Chin, L., Garraway, L. a, 2013. Highly recurrent TERT promoter mutations in human melanoma. Science 339, 957–9. doi:10.1126/science.1229259

Huang, H.W., Mullikin, J.C., Hansen, N.F., 2015. Evaluation of variant detection software for pooled next-generation sequence data. BMC Bioinformatics 16, 235. doi:10.1186/s12859-015-0624-y

Huang, M., Chen, W., Qi, F., Xia, R., Sun, M., Xu, T., Yin, L., Zhang, E., De, W., Shu, Y., 2015. Long non-coding RNA ANRIL is upregulated in hepatocellular carcinoma and regulates cell apoptosis by epigenetic silencing of KLF2. J. Hematol. Oncol. 8, 50. doi:10.1186/s13045-015-0146-0

Huang, S.-P., Lin, V.C., Lee, Y.-C., Yu, C.-C., Huang, C.-Y., Chang, T.-Y., Lee, H.-Z., Juang, S.-H., Lu, T.-L., Bao, B.-Y., 2013. Genetic variants in nuclear factor-kappa B binding sites are associated with clinical outcomes in prostate cancer patients. Eur. J. Cancer 49, 3729–37. doi:10.1016/j.ejca.2013.07.012

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., Attardi, L.D., Regev, A., Lander, E.S., Jacks, T., Rinn, J.L., 2010. A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response. Cell 142, 409–419. doi:10.1016/j.cell.2010.06.040

Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S., Guttmacher, A., Guyer, M., Hemsley, F.M., Jennings, J.L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D.P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T.S., Remacle, J., Schafer, A.J., Shibata, T., Stratton, M.R., Vockley, J.G., Watanabe, K., Yang, H., Yuen, M.M.F., Knoppers, B.M., Bobrow, M., Cambon-Thomsen, A., Dressler, L.G., Dyke, S.O.M., Joly, Y., Kato, K., Kennedy, K.L., Nicolás, P., Parker, M.J., Rial-Sebbag, E., Romeo-Casabona, C.M., Shaw, K.M., Wallace, S., Wiesner, G.L., Zeps, N., Lichter, P., Biankin, A. V, Chabannon, C., Chin, L., Clément, B., de Alava, E., Degos, F., Ferguson, M.L., Geary, P., Hayes, D.N., Hudson, T.J., Johns, A.L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. a, Sarin, R., Scarpa, A., Shibata, T., van de Vijver, M., Futreal, P.A., Aburatani, H., Bayés, M., Botwell, D.D.L., Campbell, P.J., Estivill, X., Gerhard, D.S., Grimmond, S.M., Gut, I., Hirst, M., López-Otín, C., Majumder, P., Marra, M., McPherson, J.D., Nakagawa, H., Ning, Z., Puente, X.S., Ruan, Y., Shibata, T., Stratton, M.R., Stunnenberg, H.G., Swerdlow, H., Velculescu, V.E., Wilson, R.K., Xue, H.H., Yang, L., Spellman, P.T., Bader, G.D., Boutros, P.C., Campbell, P.J., Flicek, P., Getz, G., Guigó, R., Guo, G., Haussler, D., Heath, S., Hubbard, T.J., Jiang, T., Jones, S.M., Li, Q., López-Bigas, N., Luo, R., Muthuswamy, L., Ouellette, B.F.F., Pearson, J. V, Puente, X.S., Quesada, V., Raphael, B.J., Sander, C., Shibata, T., Speed, T.P., Stein, L.D., Stuart, J.M., Teague, J.W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D. a, Wu, H., Zhao, S., Zhou, G., Stein, L.D., Guigó, R., Hubbard, T.J., Joly, Y., Jones, S.M., Kasprzyk, A., Lathrop, M., López-Bigas, N., Ouellette, B.F.F., Spellman, P.T., Teague, J.W., Thomas, G., Valencia, A., Yoshida, T., Kennedy, K.L., Axton, M., Dyke, S.O.M., Futreal, P.A., Gerhard, D.S., Gunter, C., Guyer, M., Hudson, T.J., McPherson, J.D., Miller, L.J., Ozenberger, B., Shaw, K.M., Kasprzyk, A., Stein, L.D., Zhang, J., Haider, S. a, Wang, J.,

Yung, C.K., Cros, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Bobrow, M., Chalmers, D.R.C., Hasel, K.W., Joly, Y., Kaan, T.S.H., Kennedy, K.L., Knoppers, B.M., Lowrance, W.W., Masui, T., Nicolás, P., Rial-Sebbag, E., Rodriguez, L.L., Vergely, C., Yoshida, T., Grimmond, S.M., Biankin, A. V, Bowtell, D.D.L., Cloonan, N., deFazio, A., Eshleman, J.R., Etemadmoghadam, D., Gardiner, B.B., Kench, J.G., Scarpa, A., Sutherland, R.L., Tempero, M. a, Waddell, N.J., Wilson, P.J., McPherson, J.D., Gallinger, S., Tsao, M.-S., Shaw, P. a, Petersen, G.M., Mukhopadhyay, D., Chin, L., DePinho, R. a, Thayer, S., Muthuswamy, L., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu, Y., Ji, J., Zhang, X., Chen, F., Hu, X., Zhou, G., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Yang, H., Lathrop, M., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., Egevard, L., Prokhortchouk, E., Banks, R.E., Uhlén, M., Cambon-Thomsen, A., Viksna, J., Ponten, F., Skryabin, K., Stratton, M.R., Futreal, P.A., Birney, E., Borg, A., Børresen-Dale, A.-L., Caldas, C., Foekens, J. a, Martin, S., Reis-Filho, J.S., Richardson, A.L., Sotiriou, C., Stunnenberg, H.G., Thoms, G., van de Vijver, M., van't Veer, L., Calvo, F., Birnbaum, D., Blanche, H., Boucher, P., Boyault, S., Chabannon, C., Gut, I., Masson-Jacquemier, J.D., Lathrop, M., Pauporté, I., Pivot, X., Vincent-Salomon, A., Tabone, E., Theillet, C., Thomas, G., Tost, J., Treilleux, I., Calvo, F., Bioulac-Sage, P., Clément, B., Decaens, T., Degos, F., Franco, D., Gut, I., Gut, M., Heath, S., Lathrop, M., Samuel, D., Thomas, G., Zucman-Rossi, J., Lichter, P., Eils, R., Brors, B., Korbel, J.O., Korshunov, A., Landgraf, P., Lehrach, H., Pfister, S., Radlwimmer, B., Reifenberger, G., Taylor, M.D., von Kalle, C., Majumder, P.P., Sarin, R., Rao, T.S., Bhan, M.K., Scarpa, A., Pederzoli, P., Lawlor, R. a, Delledonne, M., Bardelli, A., Biankin, A. V, Grimmond, S.M., Gress, T., Klimstra, D., Zamboni, G., Shibata, T., Nakamura, Y., Nakagawa, H., Kusada, J., Tsunoda, T., Miyano, S., Aburatani, H., Kato, K., Fujimoto, A., Yoshida, T., Campo, E., López-Otín, C., Estivill, X., Guigó, R., de Sanjosé, S., Piris, M. a, Montserrat, E., González-Díaz, M., Puente, X.S., Jares, P., Valencia, A., Himmelbauer, H., Quesada, V., Bea, S., Stratton, M.R., Futreal, P.A., Campbell, P.J., Vincent-Salomon, A., Richardson, A.L., Reis-Filho, J.S., van de Vijver, M., Thomas, G., Masson-Jacquemier, J.D., Aparicio, S., Borg, A., Børresen-Dale, A.-L., Caldas, C., Foekens, J. a, Stunnenberg, H.G., van't Veer, L., Easton, D.F., Spellman, P.T., Martin, S., Barker, A.D., Chin, L., Collins, F.S., Compton, C.C., Ferguson, M.L., Gerhard, D.S., Getz, G., Gunter, C., Guttmacher, A., Guyer, M., Hayes, D.N., Lander, E.S., Ozenberger, B., Penny, R., Peterson, J., Sander, C., Shaw, K.M., Speed, T.P., Spellman, P.T., Vockley, J.G., Wheeler, D. a, Wilson, R.K., Hudson, T.J., Chin, L., Knoppers, B.M., Lander, E.S., Lichter, P., Stein, L.D., Stratton, M.R., Anderson, W., Barker, A.D., Bell, C., Bobrow, M., Burke, W., Collins, F.S., Compton, C.C., DePinho, R. a, Easton, D.F., Futreal, P.A., Gerhard, D.S., Green, A.R., Guyer, M., Hamilton, S.R., Hubbard, T.J., Kallioniemi, O.P., Kennedy, K.L., Ley, T.J., Liu, E.T., Lu, Y., Majumder, P., Marra, M., Ozenberger, B., Peterson, J., Schafer, A.J., Spellman, P.T., Stunnenberg, H.G., Wainwright, B.J., Wilson, R.K., Yang, H., 2010. International network of cancer genome projects. Nature 464, 993–998. doi:10.1038/nature08987

Hwang, H.C., Clurman, B.E., 2005. Cyclin E in normal and neoplastic cell cycles. Oncogene 24, 2776–2786. doi:10.1038/sj.onc.1208613

Iacobucci, I., Sazzini, M., Garagnani, P., Ferrari, A., Boattini, A., Lonetti, A., Papayannidis, C., Mantovani, V., Marasco, E., Ottaviani, E., Soverini, S., Girelli, D., Luiselli, D., Vignetti, M., Baccarani, M., Martinelli, G., 2011. A polymorphism in the chromosome

9p21 ANRIL locus is associated to Philadelphia positive acute lymphoblastic leukemia. Leuk. Res. 35, 1052–9. doi:10.1016/j.leukres.2011.02.020

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S.M., Wu, Y., Robinson, D.R., Beer, D.G., Feng, F.Y., Iyer, H.K., Chinnaiyan, A.M., 2015. The landscape of long noncoding RNAs in the human transcriptome 47. doi:10.1038/ng.3192

Jeggari, A., Marks, D.S., Larsson, E., 2012. miRcode: A map of putative microrna target sites in the long non-coding transcriptome. Bioinformatics 28, 2062–2063. doi:10.1093/bioinformatics/bts344

Jeon, Y., Sarma, K., Lee, J.T., 2012. New and Xisting regulatory mechanisms of X chromosome inactivation. Curr. Opin. Genet. Dev. 22, 62–71. doi:10.1016/j.gde.2012.02.007

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W.E., Serve, H., Müller-Tidow, C., 2003. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22, 8031–8041. doi:10.1038/sj.onc.1206928

Jiang, J., Jia, P., Shen, B., Zhao, Z., 2012. Top associated SNPs in prostate cancer are significantly enriched in cis -expression quantitative trait loci and at transcription factor binding sites 5.

Jin, G., Sun, J., Isaacs, S.D., Wiley, K.E., Kim, S.T., Chu, L.W., Zhang, Z., Zhao, H., Zheng, S.L., Isaacs, W.B., Xu, J., 2011. Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. Carcinogenesis 32, 1655–1659. doi:10.1093/carcin/bgr187

Jin, H.W., Ichikawa, H., Fujita, M., Yamaai, T., Mukae, K., Nomura, K., Sugimoto, T., 2005. Involvement of caspase cascade in capsaicin-induced apoptosis of dorsal root ganglion neurons. Brain Res. 1056, 139–144. doi:10.1016/j.brainres.2005.07.025

Jolly, K.W., Malkin, D., Douglass, E.C., Brown, T.F., Sinclair, a E., Look, a T., 1994. Splice-site mutation of the p53 gene in a family with hereditary breast-ovarian cancer. Oncogene 9, 97–102.

Ju, Y.S., Lee, W.-C., Shin, J.-Y., Lee, S., Bleazard, T., Won, J.-K., Kim, Y.T., Kim, J.-I., Kang, J.-H., Seo, J.-S., 2012. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res. 22, 436–45. doi:10.1101/gr.133645.111

Kamaraj, B., Gopalakrishnan, C., Purohit, R., 2014. In Silico Analysis of miRNA-Mediated Gene Regulation in OCA and OA Genes. Cell Biochem. Biophys. 70, 12013. doi:10.1007/s12013-014-0152-9

Kaminker, J.S., Zhang, Y., Watanabe, C., Zhang, Z., 2007a. CanPredict: A computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res. 35, 595–598. doi:10.1093/nar/gkm405

Kaminker, J.S., Zhang, Y., Waugh, A., Haverty, P.M., Peters, B., Sebisanovic, D., Stinson, J., Forrest, W.F., Bazan, J.F., Seshagiri, S., Zhang, Z., 2007b. Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res. 67, 465–473. doi:10.1158/0008-5472.CAN-06-1736

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M. a, Leiserson, M.D.M., Miller, C. a, Welch, J.S., Walter, M.J., Wendl, M.C., Ley, T.J., Wilson, R.K., Raphael, B.J., Ding, L., 2013. Mutational landscape and significance across 12 major cancer types. Nature 502, 333–9. doi:10.1038/nature12634

Karolchik, D., 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, 493D–496. doi:10.1093/nar/gkh103

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P. a., Guruvadoo, L., Haeussler, M., Harte, R. a., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., Sloan, C. a., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., Kent, W.J., 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 42, 764–770. doi:10.1093/nar/gkt1168

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., Hong, M.-Y., Karczewski, K.J., Huber, W., Weissman, S.M., Gerstein, M.B., Korbel, J.O., Snyder, M., 2010. Variation in transcription factor binding among humans. Science 328, 232–235. doi:10.1126/science.1183621

Khaitan, D., Dinger, M.E., Mazar, J., Crawford, J., Smith, M. a., Mattick, J.S., Perera, R.J., 2011. The Melanoma-Upregulated Long Noncoding RNA SPRY4-IT1 Modulates Apoptosis and Invasion. Cancer Res. 71, 3852–3862. doi:10.1158/0008-5472.CAN-10-4460

Khosravi-Far, R., Esposti, M.D., 2004. Death receptor signals to mitochondria. Cancer Biol. Ther. 3, 1051–7. doi:10.4161/cbt.3.11.1173

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U.S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüs, Z.H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S.M., MacArthur, D.G., Marth, G., Muzny, D., Pers, T.H., Ritchie, G.R.S., Rosenfeld, J. a, Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., Dermitzakis, E.T., Yu, H., Rubin, M. a, Tyler-Smith, C., Gerstein, M., 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342, 1235587. doi:10.1126/science.1235587

Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L. a, Friedman, A.H., Friedman, H., Gallia, G.L., Giovanella, B.C., Grollman, A.P., He, T.-C., He, Y., Hruban, R.H., Jallo, G.I., Mandahl, N., Meeker, A.K., Mertens, F., Netto, G.J., Rasheed, B.A., Riggins, G.J., Rosenquist, T. a, Schiffman, M., Shih, I.-M., Theodorescu, D., Torbenson, M.S., Velculescu, V.E., Wang, T.-L., Wentzensen, N., Wood, L.D., Zhang, M., McLendon, R.E., Bigner, D.D., Kinzler, K.W., Vogelstein, B., Papadopoulos, N., Yan, H., 2013. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. U. S. A. 110, 6021–6. doi:10.1073/pnas.1303607110

Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., Kim, S., Safe, S., 2013. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. Oncogene 32, 1616–25. doi:10.1038/onc.2012.193

Kino, M., Hur, D.E., Ichijo, T., Nader, N., Chrousos, G.P., 2010. Noncoding RNA Gas5 Is a Growth Arrest and Starvation- Associated Repressor of the Glucocorticoid Receptor. Sci. Signal. 3. doi:10.1126/scisignal.2000568.Noncoding

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., Shendure, J., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–5. doi:10.1038/ng.2892

Kitagawa, M., Higashi, H., Jung, H.K., Suzuki-Takahashi, I., Ikeda, M., Tamai, K., Kato, J., Segawa, K., Yoshida, E., Nishimura, S., Taya, Y., 1996. The consensus motif for phosphorylation by cyclin D1-Cdk4 is different from that for phosphorylation by cyclin A/E-Cdk2. EMBO J. 15, 7060–9.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., Mclellan, M.D., Lin, L., Miller, C. a, Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing 568–576. doi:10.1101/gr.129684.111

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, a., Komune, S., Miyano, S., Mori, M., 2011. Long Noncoding RNA HOTAIR Regulates Polycomb-Dependent Chromatin Modification and Is Associated with Poor Prognosis in Colorectal Cancers. Cancer Res. 71, 6320–6326. doi:10.1158/0008-5472.CAN-11-1021

Kok, J.B. De, Verhaegh, G.W., Roelofs, R.W., Hessels, D., Kiemeney, L. a., Aalders, T.W., Swinkels, D.W., Schalken, J. a., 2002. DD3PCA3, a Very Sensitive and Specific Marker to Detect Prostate Tumors. Cancer Res. 62, 2695–2698.

Korneev, S. a, Korneeva, E.I., Lagarkova, M. a, Kiselev, S.L., Critchley, G., O'Shea, M., 2008. Novel noncoding antisense RNA transcribed from human anti-NOS2A locus is differentially regulated during neuronal differentiation of embryonic stem cells. RNA 14, 2030–2037. doi:10.1261/rna.1084308

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., Xiong, Y., 2011. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene 30, 1956–1962. doi:10.1038/onc.2010.568

Krawczak, M., Thomas, N.S.T., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., Cooper, D.N., 2007. Single base-pair substitutions in exon-intron junctions of human genes: Nature, distribution, and consequences for mRNA splicing. Hum. Mutat. 28, 150–158. doi:10.1002/humu.20400

Kurahashi, H., Takami, K., Oue, T., 1995. Biallelic Inactivation of the APC Gene in Hepatoblastoma Binlie ! ic Inactivation of the APC Gene in Hepatoblastoma1 5, 5007–5011.

Kwanhian, W., Lenze, D., Alles, J., Motsch, N., Barth, S., Döll, C., Imig, J., Hummel, M., Tinguely, M., Trivedi, P., Lulitanond, V., Meister, G., Renner, C., Grässer, F. a, 2012. MicroRNA-142 is mutated in about 20% of diffuse large B-cell lymphoma. Cancer Med. 1, 141–55. doi:10.1002/cam4.29

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., Maglott, D.R., 2014. ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42, 980–985. doi:10.1093/nar/gkt1113

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S. a, Kiezun, A., Hammerman, P.S., McKenna, A., Drier, Y., Zou, L., Ramos, A.H., Pugh, T.J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M.L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D.I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A.M., Lohr, J., Landau, D.-A., Wu, C.J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. a, Mora, J., Lee, R.S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S.B., Roberts, C.W.M., Biegel, J. a, Stegmaier, K., Bass, A.J., Garraway, L. a, Meyerson, M., Golub, T.R., Gordenin, D. a, Sunyaev, S., Lander, E.S., Getz, G., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–8. doi:10.1038/nature12213

Léveillé, N., Melo, C. a., Rooijers, K., Díaz-Lagares, A., Melo, S. a., Korkmaz, G., Lopes, R., Moqadam, F.A., Maia, A.R., Wijchers, P.J., Geeven, G., den Boer, M.L., Kalluri, R., de Laat, W., Esteller, M., Agami, R., 2015. Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Nat. Commun. 6, 6520. doi:10.1038/ncomms7520

Leygue, E., Dotzlaw, H., Watson, P.H., Murphy, L.C., 1999. Expression of the Steroid Receptor RNA Activator in Human Breast Tumors Advances in Brief Expression of the Steroid Receptor RNA Activator in Human Breast Tumors 1. Cancer Res. 59, 4190–4193.

Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., Radivojac, P., 2009. Automated inference of molecular mechanisms of disease from

amino acid substitutions. Bioinformatics 25, 2744–2750. doi:10.1093/bioinformatics/btp528

Li, L., Sun, R., Liang, Y., Pan, X., Li, Z., Bai, P., Zeng, X., Zhang, D., Zhang, L., Gao, L., 2013. Association between polymorphisms in long non-coding RNA PRNCR1 in 8q24 and risk of colorectal cancer. J. Exp. Clin. Cancer Res. 32, 1–7. doi:http://dx.doi.org/10.1186/1756-9966-32-104

Li, L.J., Zhu, J.L., Bao, W.S., Chen, D.K., Huang, W.W., Weng, Z.L., 2014. Long noncoding RNA GHET1 promotes the development of bladder cancer. Int J Clin Exp Pathol 7, 7196–7205.

Li, T., Xie, J., Shen, C., Cheng, D., Shi, Y., Wu, Z., Deng, X., Chen, H., Shen, B., Peng, C., Li, H., Zhan, Q., Zhu, Z., 2015. Upregulation of long noncoding RNA ZEB1-AS1 promotes tumor metastasis and predicts poor prognosis in hepatocellular carcinoma. Oncogene 1–10. doi:10.1038/onc.2015.223

Li, X., Wu, Z., Mei, Q., Guo, M., Fu, X., Han, W., 2013. Long non-coding RNA HOTAIR, a driver of malignancy, predicts negative prognosis and exhibits oncogenic activity in oesophageal squamous cell carcinoma. Br. J. Cancer 109, 2266–78. doi:10.1038/bjc.2013.548

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D., Zhao, H., Skogerbø, G., Wu, Z., Zhao, Y., 2011. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res. 39, 3864–78. doi:10.1093/nar/gkq1348

Liebert, M.A., Gene, P.C.N., 2006. Regulation of Apoptosis by a Prostate-Specific and. DNA Cell Biol. 25, 135–141.

Lin, R., Maeda, S., Liu, C., Karin, M., Edgington, T.S., 2007. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. Oncogene 26, 851–8. doi:10.1038/sj.onc.1209846

Lin, V.C., Huang, C.-Y., Lee, Y.-C., Yu, C.-C., Chang, T.-Y., Lu, T.-L., Huang, S.-P., Bao, B.-Y., 2014. Genetic variations in TP53 binding sites are predictors of clinical outcomes in prostate cancer patients. Arch. Toxicol. 88, 901–11. doi:10.1007/s00204-014-1196-8

Ling, H., Spizzo, R., Atlasi, Y., Nicoloso, M., Shimizu, M., Redis, R.S., Nishida, N., Gafà, R., Song, J., Guo, Z., Ivan, C., Barbarotto, E., De Vries, I., Zhang, X., Ferracin, M., Churchman, M., Van Galen, J.F., Beverloo, B.H., Shariati, M., Haderk, F., Estecio, M.R., Garcia-Manero, G., Patijn, G. a., Gotley, D.C., Bhardwaj, V., Shureiqi, I., Sen, S., Multani, A.S., Welsh, J., Yamamoto, K., Taniguchi, I., Song, M.A., Gallinger, S., Casey, G., Thibodeau, S.N., Le Marchand, L., Tiirikainen, M., Mani, S. a., Zhang, W., Davuluri, R. V., Mimori, K., Mori, M., Sieuwerts, A.M., Martens, J.W.M., Tomlinson, I., Negrini, M., Berindan-Neagoe, I., Foekens, J. a., Hamilton, S.R., Lanza, G., Kopetz, S., Fodde, R., Calin, G. a., 2013. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. Genome Res. 23, 1446–1461. doi:10.1101/gr.152942.112

Liu, M.-X., Chen, X., Chen, G., Cui, Q.-H., Yan, G.-Y., 2014. A Computational Framework to Infer Human Disease-Associated Long Noncoding RNAs. PLoS One 9, e84408. doi:10.1371/journal.pone.0084408

Liu, Q., Huang, J., Zhou, N., Zhang, Z., Zhang, a., Lu, Z., Wu, F., Mo, Y.-Y., 2013. LncRNA loc285194 is a p53-regulated tumor suppressor. Nucleic Acids Res. 41, 4976–4987. doi:10.1093/nar/gkt182

Liu, Q., Huang, J., Zhou, N., Zhang, Z., Zhang, A., Lu, Z., Wu, F., Mo, Y.Y., 2013. LncRNA loc285194 is a p53-regulated tumor suppressor. Nucleic Acids Res. 41, 4976–4987. doi:10.1093/nar/gkt182

Liu, X., Li, D., Zhang, W., Guo, M., Zhan, Q., 2012. Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. EMBO J. 31, 4415–27. doi:10.1038/emboj.2012.292

Liu, Y., Sharma, S., Watabe, K., 2015. Roles of lncRNA in breast cancer. Front. Biosci. (Schol. Ed). 7, 94–108.

Liu, Z., Wang, W., Jiang, J., Bao, E., Xu, D., Zeng, Y., Tao, L., Qiu, J., 2013. Downregulation of GAS5 promotes bladder cancer cell proliferation, partly by regulating CDK6. PLoS One 8, e73991. doi:10.1371/journal.pone.0073991

Lomelin, D., Jorgenson, E., Risch, N., 2010. Human genetic variation recognizes functional elements in noncoding sequence Human genetic variation recognizes functional elements in noncoding sequence. Genome Res 20, 311–319. doi:10.1101/gr.094151.109

Lopes, M.C., Joyce, C., Ritchie, G.R.S., John, S.L., Cunningham, F., Asimit, J., Zeggini, E., 2012. A combined functional annotation score for non-synonymous variants. Hum. Hered. 73, 47–51. doi:10.1159/000334984

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. bioRxiv 1–21. doi:10.1101/002832

Lowe, C.B., Haussler, D., 2012. 29 Mammalian Genomes Reveal Novel Exaptations of Mobile Elements for Likely Regulatory Functions in the Human Genome. PLoS One 7. doi:10.1371/journal.pone.0043128

Lu, K., Li, W., Liu, X., Sun, M., Zhang, M., Wu, W., Xie, W., Hou, Y., 2013. Long non-coding RNA MEG3 inhibits NSCLC cells proliferation and induces apoptosis by affecting p53 expression. BMC Cancer 13, 461. doi:10.1186/1471-2407-13-461

Luo, M., Li, Z., Wang, W., Zeng, Y., Liu, Z., Qiu, J., 2013. Long non-coding RNA H19 increases bladder cancer metastasis by associating with EZH2 and inhibiting E-cadherin expression. Cancer Lett. 333, 213–221. doi:10.1016/j.canlet.2013.01.033

Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J., Tian, W., 2012. Molecular mechanisms and function prediction of long noncoding RNA. ScientificWorldJournal. 2012, 541786. doi:10.1100/2012/541786

MacArthur, D.G., Manolio, T. a, Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E. a, Barrett, J.C., Biesecker, L.G., Conrad, D.F., Cooper, G.M., Cox, N.J., Daly, M.J., Gerstein, M.B., Goldstein, D.B., Hirschhorn, J.N., Leal, S.M., Pennacchio, L. a, Stamatoyannopoulos, J. a, Sunyaev, S.R., Valle, D., Voight, B.F., Winckler, W., Gunter, C., 2014. Guidelines for investigating causality of sequence variants in human disease. Nature 508, 469–76. doi:10.1038/nature13127

Manikandan, M., Munirajan, A.K., 2014. Single nucleotide polymorphisms in microRNA binding sites of oncogenes: implications in cancer and pharmacogenomics. Omi. a J. Integr. Biol. 18, 142–54. doi:10.1089/omi.2013.0098

Manikandan, M., Raksha, G., Munirajan, A.K., 2012. Haploinsufficiency of tumor suppressor genes is driven by the cumulative effect of micrornas, microrna binding site polymorphisms and microrna polymorphisms: An in silico approach. Cancer Inform. 11, 157–171. doi:10.4137/CIN.S10176

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., Akoulitchev, A., 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445, 666–670. doi:10.1038/nature05519

Matouk, I.J., DeGroot, N., Mezan, S., Ayesh, S., Abu-lail, R., Hochberg, A., Galun, E., 2007. The H19 Non-Coding RNA Is Essential for Human Tumor Growth. PLoS One 2, e845. doi:10.1371/journal.pone.0000845

Matouk, I.J., Raveh, E., Abu-lail, R., Mezan, S., Gilon, M., Gershtain, E., Birman, T., Gallula, J., Schneider, T., Barkali, M., Richler, C., Fellig, Y., Sorin, V., Hubert, A., Hochberg, A., Czerniak, A., 2014. Oncofetal H19 RNA promotes tumor metastasis. Biochim. Biophys. Acta - Mol. Cell Res. 1843, 1414–1426. doi:10.1016/j.bbamcr.2014.03.023

Mattick, J.S., Amaral, P.P., Dinger, M.E., Mercer, T.R., Mehler, M.F., 2009. RNA regulation of epigenetic processes. BioEssays 31, 51–59. doi:10.1002/bies.080099

Mcdaniell, R., Lee, B., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., Keefe, D., Collins, F.S., Willard, H.F., Lieb, J.D., Furey, T.S., Crawford, G.E., Iyer, V.R., Birney, E., 2010. Heritable Individual-Specific 328, 235–240.

McHugh, C. a., Chen, C.-K., Chow, A., Surka, C.F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M.J., Shishkin, A. a., Su, J., Lander, E.S., Hess, S., Plath, K., Guttman, M., 2015. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature. doi:10.1038/nature14443

Moreau, Y., Tranchevent, L.-C., 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat. Rev. Genet. 13, 523–536. doi:10.1038/nrg3253

Morgan, D.O., 1995. Principles of CDK regulation. Nature. doi:10.1038/374131a0

Negishi, M., Wongpalee, S.P., Sarkar, S., Park, J., Lee, K.Y., Shibata, Y., Reon, B.J., Abounader, R., Suzuki, Y., Sugano, S., Dutta, A., 2014. A New lncRNA, APTR, Associates with and Represses the CDKN1A/p21 Promoter by Recruiting Polycomb Proteins. PLoS One 9, e95216. doi:10.1371/journal.pone.0095216

Ng, P.C., Henikoff, S., 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814. doi:10.1093/nar/gkg509

Nie, L., Wu, H.J., Hsu, J.M., Chang, S.S., Labaff, a M., Li, C.W., Wang, Y., Hsu, J.L., Hung, M.C., 2012. Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer. Am J Transl Res 4, 127–150.

Niinuma, T., Suzuki, H., Nojima, M., Nosho, K., Yamamoto, H., Takamaru, H., Yamamoto, E., Maruyama, R., Nobuoka, T., Miyazaki, Y., Nishida, T., Bamba, T., Kanda, T., Ajioka, Y., Taguchi, T., Okahara, S., Takahashi, H., Nishida, Y., Hosokawa, M., Hasegawa, T., Tokino, T., Hirata, K., Imai, K., Toyota, M., Shinomura, Y., 2012. Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors. Cancer Res. 72, 1126–36. doi:10.1158/0008-5472.CAN-11-1803

Nitsche A, Rose D, Fasold M, Reiche K, S.P., 2015. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved . 21, 46342. doi:10.1261/rna.046342.114.

Okugawa, Y., Toiyama, Y., Hur, K., Toden, S., Saigusa, S., Tanaka, K., Inoue, Y., Mohri, Y., Kusunoki, M., Cr, B., Goel, A., 2014. Metastasis-associated long non-coding RNA drives gastric cancer development and promotes peritoneal metastasis . 35, 2731. doi:10.1093/carcin/bgu200.

Orom, U. a., Derrien, T., Guigo, R., Shiekhattar, R., 2010. Long Noncoding RNAs as Enhancers of Gene Expression. Cold Spring Harb. Symp. Quant. Biol. 75, 325–331. doi:10.1101/sqb.2010.75.058

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z., 2014. A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. 15, 256–278. doi:10.1093/bib/bbs086

Palii, C.G., Perez-Iratxeta, C., Yao, Z., Cao, Y., Dai, F., Davison, J., Atkins, H., Allan, D., Dilworth, F.J., Gentleman, R., Tapscott, S.J., Brand, M., 2011. Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. EMBO J. 30, 494–509. doi:10.1038/emboj.2010.342

Panzitt, K., Tschernatsch, M.M.O., Guelly, C., Moustafa, T., Stradner, M., Strohmaier, H.M., Buck, C.R., Denk, H., Schroeder, R., Trauner, M., Zatloukal, K., 2007. Characterization of HULC, a Novel Gene With Striking Up-Regulation in Hepatocellular Carcinoma, as Noncoding RNA. Gastroenterology 132, 330–342. doi:10.1053/j.gastro.2006.08.026

Pasic, I., Shlien, A., Durbin, A.D., Stavropoulos, D.J., Baskin, B., Ray, P.N., Novokmet, A., Malkin, D., 2010. Recurrent focal copy-number changes and loss of heterozygosity

implicate two noncoding RNAs and one tumor suppressor gene at chromosome 3q13.31 in osteosarcoma. Cancer Res. 70, 160–171. doi:10.1158/0008-5472.CAN-09-1902

Peinado, H., Olmeda, D., Cano, A., 2007. Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? Nat. Rev. Cancer 7, 415–428. doi:10.1038/nrc2131

Petrovics, G., Zhang, W., Makarem, M., Street, J.P., Connelly, R., Sun, L., Sesterhenn, I. a, Srikantan, V., Moul, J.W., Srivastava, S., 2004. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. Oncogene 23, 605–11. doi:10.1038/sj.onc.1207069

Pj, B., Cao, H., Np, C., Gk, C., Davis, S., Day, N., Dhami, P., Sc, D., Fiegler, H., Pg, G., Goldy, J., Hawrylycz, M., Haydock, a, Humbert, R., Kd, J., Em, J., Tt, F., Er, R., Karnani, N., Lee, K., Gc, L., Pa, N., Sc, P., Pj, S., Sandstrom, R., Shafer, a, Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Reymond, a, Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, a, Harrow, J., Zhao, X., Kg, S., Wk, S., Hs, O., Kp, C., Foissac, S., Alioto, T., Brent, M., Pachter, L., 2012. Mohammad Heydarian thesis proposal - 1.19.2012 Page 27. Genes … 447, 27–32. doi:10.1101/gad.17446611.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.-L., Ordóñez, G.R., Bignell, G.R., Ye, K., Alipaz, J., Bauer, M.J., Beare, D., Butler, A., Carter, R.J., Chen, L., Cox, A.J., Edkins, S., Kokko-Gonzales, P.I., Gormley, N. a, Grocock, R.J., Haudenschild, C.D., Hims, M.M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L.J., Ning, Z., Royce, T., Schulz-Trieglaff, O.B., Spiridou, A., Stebbings, L. a, Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M.T., Campbell, P.J., Bentley, D.R., Futreal, P.A., Stratton, M.R., 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463, 191–196. doi:10.1038/nature08658

Podlaha, O., Riester, M., De, S., Michor, F., 2012. Evolution of the cancer genome. Trends Genet. 28, 155–163. doi:10.1016/j.tig.2012.01.003

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., Pandolfi, P.P., 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465, 1033–1038. doi:10.1038/nature09144

Polyak, K., Weinberg, R. a., 2009. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat. Rev. Cancer 9, 265–273. doi:10.1038/nrc2620

Ponting, C.P., Hardison, R.C., 2011. What fraction of the human genome is functional? Genome Res. 21, 1769–1776. doi:10.1101/gr.116814.110

Ponting, C.P., Oliver, P.L., Reik, W., 2009. Evolution and Functions of Long Noncoding RNAs. Cell 136, 629–641. doi:10.1016/j.cell.2009.02.006

Prensner, J.R., Chinnaiyan, A.M., 2011. The emergence of lncRNAs in cancer biology. Cancer Discov. 1, 391–407. doi:10.1158/2159-8290.CD-11-0209

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I. a, Grasso, C.S., Kominsky, H.D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J.T., Robinson, D., Iyer, H.K., Palanisamy, N., Maher, C. a, Chinnaiyan, A.M., 2011. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. 29, 742–749. doi:10.1038/nbt.1914

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65. doi:10.1093/nar/gkl842

Quagliata, L., Matter, M.S., Piscuoglio, S., Arabi, L., Ruiz, C., Procino, A., Kovac, M., Moretti, F., Makowska, Z., Boldanova, T., Andersen, J.B., Hämmerle, M., Tornillo, L., Heim, M.H., Diederichs, S., Cillo, C., Terracciano, L.M., 2014. Long noncoding RNA HOTTIP/HOXA13 expression is associated with disease progression and predicts outcome in hepatocellular carcinoma patients. Hepatology 59, 911–23. doi:10.1002/hep.26740

Quinlan, A.R., Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. doi:10.1093/bioinformatics/btq033

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. a, Flynn, R. a, Wysocka, J., 2011. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279–283. doi:10.1038/nature09692

Rando, T. a., Chang, H.Y., 2012. Aging, rejuvenation, and epigenetic reprogramming: Resetting the aging clock. Cell 148, 46–57. doi:10.1016/j.cell.2012.01.003

Ren, S., Peng, Z., Mao, J.-H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., Chen, C., Wang, F., Guo, X., Lu, J., Yang, J., Wei, M., Tian, Z., Guan, Y., Tang, L., Xu, C., Wang, L., Gao, X., Tian, W., Wang, J., Yang, H., Wang, J., Sun, Y., 2012. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res. 22, 806–821. doi:10.1038/cr.2012.30

Reva, B., Antipin, Y., Sander, C., 2011. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. 39, 37–43. doi:10.1093/nar/gkr407

Rhee, H.S., Pugh, B.F., 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147, 1408–1419. doi:10.1016/j.cell.2011.11.013

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S. a, Goodnough, L.H., Helms, J. a, Farnham, P.J., Segal, E., Chang, H.Y., 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311–23. doi:10.1016/j.cell.2007.05.022

Rintala-Maki, N.D., Sutherland, L.C., 2009. Identification and characterisation of a novel antisense non-coding RNA from the RBM5 gene locus. Gene 445, 7–16. doi:10.1016/j.gene.2009.06.009

Ritchie, G.R.S., Dunham, I., Zeggini, E., Flicek, P., 2014. Functional annotation of noncoding sequence variants. Nat. Methods 11, 294–6. doi:10.1038/nmeth.2832

Rosenbloom, K.R., Sloan, C. a., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., Lee, B.T., Barber, G.P., Harte, R. a., Diekhans, M., Long, J.C., Wilder, S.P., Zweig, A.S., Karolchik, D., Kuhn, R.M., Haussler, D., Kent, W.J., 2013. ENCODE Data in the UCSC Genome Browser: Year 5 update. Nucleic Acids Res. 41, 56–63. doi:10.1093/nar/gks1172

Sakurai, K., Reon, B.J., Anaya, J., Dutta, a., 2015. The lncRNA DRAIC/PCAT29 Locus Constitutes a Tumor-Suppressive Nexus. Mol. Cancer Res. 13, 828–838. doi:10.1158/1541-7786.MCR-15-0016-T

Sánchez, Y., Segura, V., Marín-Béjar, O., Athie, A., Marchese, F.P., González, J., Bujanda, L., Guo, S., Matheu, A., Huarte, M., 2014. Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. Nat. Commun. 5, 5812. doi:10.1038/ncomms6812

Sato, Y., Yoshizato, T., Shiraishi, Y., Maekawa, S., Okuno, Y., Kamura, T., Shimamura, T., Sato-Otsubo, A., Nagae, G., Suzuki, H., Nagata, Y., Yoshida, K., Kon, A., Suzuki, Y., Chiba, K., Tanaka, H., Niida, A., Fujimoto, A., Tsunoda, T., Morikawa, T., Maeda, D., Kume, H., Sugano, S., Fukayama, M., Aburatani, H., Sanada, M., Miyano, S., Homma, Y., Ogawa, S., 2013. Integrated molecular analysis of clear-cell renal cell carcinoma. Nat. Genet. 45, 860–7. doi:10.1038/ng.2699

Schmidt, L.H., Spieker, T., Koschmieder, S., Schäffers, S., Humberg, J., Jungen, D., Bulk, E., Hascher, A., Wittmer, D., Marra, A., Hillejan, L., Wiebe, K., Berdel, W.E., Wiewrodt, R., Muller-Tidow, C., 2011. The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. J. Thorac. Oncol. 6, 1984–92. doi:10.1097/JTO.0b013e3182307eac

Schneider, C., King, R.M., Philipson, L., 1988. Genes specifically expressed at growth arrest of mammalian cells. Cell 54, 787–793. doi:10.1016/S0092-8674(88)91065-3

Schuster-Böckler, B., Lehner, B., 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488, 504–507. doi:10.1038/nature11273

Schwartz, D., Rotter, V., 1998. P53-Dependent Cell Cycle Control: Response To Genotoxic Stress. Semin. Cancer Biol. 8, 325–336.

Shapiro, I.M., Cheng, A.W., Flytzanis, N.C., Balsamo, M., Condeelis, J.S., Oktay, M.H., Burge, C.B., Gertler, F.B., 2011. An emt-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. PLoS Genet. 7. doi:10.1371/journal.pgen.1002218

Sherr, C.J., Roberts, J.M., 1999. CDK inhibitors: positive and negative regulators of G1-phase progression. Genes Dev. 13, 1501–1512. doi:10.1101/gad.13.12.1501

Shi, S.-J., Wang, L.-J., Yu, B., Li, Y.-H., Jin, Y., Bai, X.-Z., 2015. LncRNA-ATB promotes trastuzumab resistance and invasion-metastasis cascade in breast cancer. Oncotarget 6, 11652–63.

Shi, X., Sun, M., Liu, H., Yao, Y., Kong, R., Chen, F., Song, Y., 2013. MOLECULAR CARCINOGENESIS A Critical Role for the Long Non-Coding RNA GAS5 in Proliferation and Apoptosis in Non-Small-Cell Lung Cancer 1–12. doi:10.1002/mc.22120

Shihab, H. a., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., Campbell, C., 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31, 1536–1543. doi:10.1093/bioinformatics/btv009

Shintani, Y., Fukumoto, Y., Chaika, N., Svoboda, R., Wheelock, M.J., Johnson, K.R., 2008. Collagen I-mediated up-regulation of N-cadherin requires cooperative signals from integrins and discoidin domain receptor 1. J. Cell Biol. 180, 1277–1289. doi:10.1083/jcb.200708137

Sjöblom, T., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S.D., Willis, J., Dawson, D., Willson, J.K. V, Gazdar, A.F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B.H., Bachman, K.E., 2006. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. Science (80-.). 314, 268–274. doi:10.1126/science.1133427

Smith, M. a., Gesell, T., Stadler, P.F., Mattick, J.S., 2013. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res. 41, 8220–8236. doi:10.1093/nar/gkt596

Srikantan, V., Zou, Z., Petrovics, G., Xu, L., Augustus, M., Davis, L., Livezey, J.R., Connell, T., Sesterhenn, I. a, Yoshino, K., Buzard, G.S., Mostofi, F.K., McLeod, D.G., Moul, J.W., Srivastava, S., 2000. PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. Proc. Natl. Acad. Sci. U. S. A. 97, 12216–21. doi:10.1073/pnas.97.22.12216

Stenson, P.D., Mort, M., Ball, E. V, Howells, K., Phillips, A.D., Thomas, N.S., Cooper, D.N., 2009. The Human Gene Mutation Database: 2008 update. Genome Med. 1, 13. doi:10.1186/gm13

Sterne-Weiler, T., Sanford, J.R., 2014. Exon identity crisis: disease-causing mutations that disrupt the splicing code. Genome Biol. 15, 201. doi:10.1186/gb4150

Stone, E. a, Sidow, A., 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity 978–986. doi:10.1101/gr.3804205

Stratton, M.R., Campbell, P.J., Futreal, P.A., 2009. The cancer genome. Nature 458, 719–724. doi:10.1038/nature07943

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. a, Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R., Hogenesch, J.B., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. U. S. A. 101, 6062–7. doi:10.1073/pnas.0400782101

Sugimachi, K., Niida, A., Yamamoto, K., Shimamura, T., Imoto, S., Iinuma, H., Shinden, Y., Eguchi, H., 2014. Allelic imbalance at an 8q24 oncogenic SNP is involved in activating MYC in human colorectal cancer . 10434. doi:10.1245/s10434-013-3468-6.

Sun, M., Gadad, S.S., Kim, D.-S., Kraus, W.L., 2015. Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. Mol. Cell 1–14. doi:10.1016/j.molcel.2015.06.023

Sun, M., Jin, F., Xia, R., Kong, R., Li, J., Xu, T., Liu, Y., Zhang, E., Liu, X., De, W., 2014. Decreased expression of long noncoding RNA GAS5 indicates a poor prognosis and promotes cell proliferation in gastric cancer. BMC Cancer 14, 319. doi:10.1186/1471-2407-14-319

Sun, N., Ye, C., Zhao, Q., Zhang, Q., Xu, C., Wang, S., Jin, Z., Sun, S., Wang, F., Li, W., 2014. Long Noncoding RNA-EBIC Promotes Tumor Cell Invasion by Binding to EZH2 and Repressing E-Cadherin in Cervical Cancer. PLoS One 9, e100340. doi:10.1371/journal.pone.0100340

Tamborero, D., Gonzalez-Perez, A., Lopez-Bigas, N., 2013. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29, 2238–2244. doi:10.1093/bioinformatics/btt395

Tano, K., Mizuno, R., Okada, T., Rakwal, R., Shibato, J., Masuo, Y., Ijiri, K., Akimitsu, N., 2010. MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. FEBS Lett. 584, 4575–4580. doi:10.1016/j.febslet.2010.10.008

Taylor, M.D., Gokgoz, N., Andrulis, I.L., Mainprize, T.G., Drake, J.M., Rutka, J.T., 2000. Familial posterior fossa brain tumors of infancy secondary to germline mutation of the hSNF5 gene. Am. J. Hum. Genet. 66, 1403–1406. doi:10.1086/302833

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A., 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. 13, 2129–2141. doi:10.1101/gr.772403

Thusberg, J., Olatubosun, A., Vihinen, M., 2011. Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. 32, 358–368. doi:10.1002/humu.21445

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P. a., Blencowe, B.J., Prasanth, S.G., Prasanth, K. V., 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating

SR splicing factor phosphorylation. Mol. Cell 39, 925–938. doi:10.1016/j.molcel.2010.08.011

Tripathi, V., Shen, Z., Chakraborty, A., Giri, S., Freier, S.M., Wu, X., Zhang, Y., Gorospe, M., Prasanth, S.G., Lal, A., Prasanth, K. V., 2013. Long Noncoding RNA MALAT1 Controls Cell Cycle Progression by Regulating the Expression of Oncogenic Transcription Factor B-MYB. PLoS Genet. 9, e1003368. doi:10.1371/journal.pgen.1003368

Tseng, Y.-Y., Moriarity, B.S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., Ronning, P., Reuland, B., Guenther, K., Beadnell, T.C., Essig, J., Otto, G.M., O'Sullivan, M.G., Largaespada, D. a., Schwertfeger, K.L., Marahrens, Y., Kawakami, Y., Bagchi, A., 2014. PVT1 dependence in cancer with MYC copy-number increase. Nature 82. doi:10.1038/nature13311

Ulitsky, I., Bartel, D.P., 2013. XLincRNAs: Genomics, evolution, and mechanisms. Cell 154, 26–46. doi:10.1016/j.cell.2013.06.020

Urban, T.J., 2005. Functional genomics of membrane transporters in human populations. Genome Res. 16, 223–230. doi:10.1101/gr.4356206

Vaishnavi, V., Manikandan, M., Munirajan, A.K., 2014. Mining the 3'UTR of Autism-implicated Genes for SNPs Perturbing MicroRNA Regulation. Genomics, Proteomics Bioinforma. 12, 92–104. doi:10.1016/j.gpb.2014.01.003

Venables, J.P., Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Koh, C., Gervais-Bird, J., Lapointe, E., Froehlich, U., Durand, M., Gendron, D., Brosseau, J.P., Thibault, P., Lucier, J.F., Tremblay, K., Prinos, P., Wellinger, R.J., Chabot, B., Rancourt, C., Elela, S.A., 2008. Identification of alternative splicing markers for breast cancer. Cancer Res. 68, 9525–9531. doi:10.1158/0008-5472.CAN-08-1769

Vidal, A., Koff, A., 2000. Cell-cycle inhibitors: Three families united by a common cause. Gene 247, 1–15. doi:10.1016/S0378-1119(00)00092-5

Vitkup, D., Sander, C., Church, G.M., 2003. The amino-acid mutational spectrum of human genetic disease. Genome Biol. 4, R72. doi:10.1186/gb-2003-4-11-r72

Wang, C.-M., Wu, Q.-Q., Li, S.-Q., Chen, F.-J., Tuo, L., Xie, H.-W., Tong, Y.-S., Ji, L., Zhou, G.-Z., Cao, G., Wu, M., Lv, J., Shi, W.-H., Cao, X.-F., 2014. Upregulation of the long non-coding RNA PlncRNA-1 promotes esophageal squamous carcinoma cell proliferation and correlates with advanced clinical stage. Dig. Dis. Sci. 59, 591–7. doi:10.1007/s10620-013-2956-7

Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F., Fan, Q., 2010. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. Nucleic Acids Res. 38, 5366–5383. doi:10.1093/nar/gkq285

Wang, K., Kan, J., Yuen, S.T., Shi, S.T., Chu, K.M., Law, S., Chan, T.L., Kan, Z., Chan, A.S.Y., Tsui, W.Y., Lee, S.P., Ho, S.L., Chan, A.K.W., Cheng, G.H.W., Roberts, P.C., Rejto, P. a, Gibson, N.W., Pocalyko, D.J., Mao, M., Xu, J., Leung, S.Y., 2011. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat. Genet. 43, 1219–1223. doi:10.1038/ng.982

Wang, T., Lin, Y., Chen, Y., Yeh, C., Huang, Y., Hsieh, T., Shieh, T., Hsueh, C., Chen, T., 2015. Long non-coding RNA AOC4P suppresses hepatocellular carcinoma metastasis by enhancing vimentin degradation and inhibiting epithelial-mesenchymal transition. Oncotarget 23–30.

Wapinski, O., Chang, H.Y., 2011. Long noncoding RNAs and human disease. Trends Cell Biol. 21, 354–361. doi:10.1016/j.tcb.2011.04.001

Ward, L.D., Kellis, M., 2012. Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. 30, 1095–106. doi:10.1038/nbt.2422

Watson, I.R., Takahashi, K., Futreal, P.A., Chin, L., 2013. Emerging patterns of somatic mutations in cancer. Nat. Rev. Genet. 14, 703–18. doi:10.1038/nrg3539

Wegert, J., Ishaque, N., Vardapour, R., Geörg, C., Gu, Z., Bieg, M., Ziegler, B., Bausenwein, S., Nourkami, N., Ludwig, N., Keller, A., Grimm, C., Kneitz, S., Williams, R.D., Chagtai, T., Pritchard-Jones, K., van Sluis, P., Volckmann, R., Koster, J., Versteeg, R., Acha, T., O'Sullivan, M.J., Bode, P.K., Niggli, F., Tytgat, G.A., van Tinteren, H., van den Heuvel-Eibrink, M.M., Meese, E., Vokuhl, C., Leuschner, I., Graf, N., Eils, R., Pfister, S.M., Kool, M., Gessler, M., 2015. Mutations in the SIX1/2 Pathway and the DROSHA/DGCR8 miRNA Microprocessor Complex Underlie High-Risk Blastemal Type Wilms Tumors. Cancer Cell 27, 298–311. doi:10.1016/j.ccell.2015.01.002

Wei, X., Walia, V., Lin, J.C., Teer, J.K., Prickett, T.D., Gartner, J., Davis, S., Stemke-Hale, K., Davies, M. a, Gershenwald, J.E., Robinson, W., Robinson, S., Rosenberg, S. a, Samuels, Y., 2011. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat. Genet. 43, 442–446. doi:10.1038/ng.810

Wei, Z., Wang, W., Hu, P., Lyon, G.J., Hakonarson, H., 2011. SNVer: A statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 39, 1–13. doi:10.1093/nar/gkr599

Weinberg, R. a, 1995. The retinoblastoma protein and cell cycle control. Cell 81, 323–330. doi:10.1016/0092-8674(95)90385-2

Whitlock, M.C., 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J. Evol. Biol. 18, 1368–1373. doi:10.1111/j.1420-9101.2005.00917.x

Willingham, a T., Orth, a P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., Schultz, P.G., 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science 309, 1570–1573. doi:10.1126/science.1115901

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40, 11189–11201. doi:10.1093/nar/gks918

Woo, Y.H., Li, W.-H., 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat. Commun. 3, 1004. doi:10.1038/ncomms1982

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. a, Kaminker, J.S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J.K. V, Sukumar, S., Polyak, K., Park, B.H., Pethiyagoda, C.L., Pant, P.V.K., Ballinger, D.G., Sparks, A.B., Hartigan, J., Smith, D.R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S.D., Parmigiani, G., Kinzler, K.W., Velculescu, V.E., Vogelstein, B., 2007. The genomic landscapes of human breast and colorectal cancers. Science 318, 1108–1113. doi:10.1126/science.1145720

Wu, T., 2006. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res. 34, D150–D152. doi:10.1093/nar/gkj025

Wu, Y., Liu, H., Shi, X., Yao, Y., Yang, W., Song, Y., 2015. The long non-coding RNA HNF1A-AS1 regulates proliferation and metastasis in lung adenocarcinoma. Oncotarget 6.

Xie, B., Ding, Q., Han, H., Wu, D., 2013. MiRCancer: A microRNA-cancer association database constructed by text mining on literature. Bioinformatics 29, 638–644. doi:10.1093/bioinformatics/btt014

Xie, M., Sun, M., Zhu, Y., Xia, R., Liu, Y., Ding, J., 2015. Long noncoding RNA HOXA-AS2 promotes gastric cancer proliferation by epigenetically silencing P21 / PLK3 / DDIT3 expression 6.

Yang, C., Li, X., Wang, Y., Zhao, L., Chen, W., 2012. Long non-coding RNA UCA1 regulated cell cycle distribution via CREB through PI3-K dependent pathway in bladder carcinoma cells. Gene 496, 8–16. doi:10.1016/j.gene.2012.01.012

Yang, F., Bi, J., Xue, X., Zheng, L., Zhi, K., Hua, J., Fang, G., 2012. Up-regulated long non-coding RNA H19 contributes to proliferation of gastric cancer cells. FEBS J. 279, 3159–3165. doi:10.1111/j.1742-4658.2012.08694.x

Yang, F., Xue, X., Zheng, L., Bi, J., Zhou, Y., Zhi, K., Gu, Y., Fang, G., 2014. Long non-coding RNA GHET1 promotes gastric carcinoma cell proliferation by increasing c-Myc mRNA stability. FEBS J. 281, 802–813. doi:10.1111/febs.12625

Yang, F., Zhang, L., Huo, X., Yuan, J., Xu, D., Yuan, S., Zhu, N., Zhou, W., Yang, G., Wang, Y., Shang, J., Gao, C., Zhang, F., Wang, F., Sun, S., 2011. Long noncoding RNA high expression in hepatocellular carcinoma facilitates tumor growth through enhancer of zeste homolog 2 in humans. Hepatology 54, 1679–1689. doi:10.1002/hep.24563

Yang, H., Zhong, Y., Xie, H., Lai, X., Xu, M., Nie, Y., Liu, S., Wan, Y.-J.Y., 2013. Induction of the liver cancer-down-regulated long noncoding RNA uc002mbe.2 mediates trichostatin-induced apoptosis of liver cancer cells. Biochem. Pharmacol. 85, 1761–1769. doi:10.1016/j.bcp.2013.04.020

Yang, L., Lin, C., Jin, C., Yang, J.C., Tanasa, B., Li, W., Merkurjev, D., Ohgi, K. a, Meng, D., Zhang, J., Evans, C.P., Rosenfeld, M.G., 2013. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature 500, 598–602. doi:10.1038/nature12451

Yang, X., Song, J.H., Cheng, Y., Wu, W., Bhagat, T., Yu, Y., Abraham, J.M., Ibrahim, S., Ravich, W., Roland, B.C., Khashab, M., Singh, V.K., Shin, E.J., Yang, X., Verma, a. K., Meltzer, S.J., Mori, Y., 2014. Long non-coding RNA HNF1A-AS1 regulates proliferation and migration in oesophageal adenocarcinoma cells. Gut 63, 881–890. doi:10.1136/gutjnl-2013-305266

Yang, Y., Li, H., Hou, S., Hu, B., Liu, J., Wang, J., 2013. The Noncoding RNA Expression Profile and the Effect of lncRNA AK126698 on Cisplatin Resistance in Non-Small-Cell Lung Cancer Cell. PLoS One 8, e65309. doi:10.1371/journal.pone.0065309

Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., Zhou, M.-M., 2010. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Mol. Cell 38, 662–674. doi:10.1016/j.molcel.2010.03.021

Yekutieli, D., Benjamini, Y., 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plan. Inference 82, 171–196. doi:10.1016/S0378-3758(99)00041-5

Yendamuri, S., Trapasso, F., Ferracin, M., Cesari, R., Sevignani, C., Shimizu, M., Rattan, S., Kuroki, T., Dumon, K.R., Bullrich, F., Liu, C., Negrini, M., Williams, N.N., Kaiser, L.R., Croce, C.M., Calin, G. a, 2007. Tumor suppressor functions of ARLTS1 in lung cancers. Cancer Res. 67, 7738–7745. doi:10.1158/0008-5472.CAN-07-1481

Yilmaz, M., Christofori, G., 2009. EMT, the cytoskeleton, and cancer cell invasion. Cancer Metastasis Rev. 28, 15–33. doi:10.1007/s10555-008-9169-0

Ying, L., Chen, Q., Wang, Y., Zhou, Z., Huang, Y., Qiu, F., 2012. Upregulated MALAT-1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition. Mol. Biosyst. 8, 2289. doi:10.1039/c2mb25070e

Yu, S., Cui, K., Jothi, R., Zhao, D.M., Jing, X., Zhao, K., Xue, H.H., 2011. GABP controls a critical transcription regulatory module that is essential for maintenance and differentiation of hematopoietic stem/progenitor cells. Blood 117, 2166–2178. doi:10.1182/blood-2010-09-306563

Yuan, S.-X., Tao, Q.-F., Wang, J., Yang, F., Liu, L., Wang, L.-L., Zhang, J., Yang, Y., Liu, H., Wang, F., Sun, S.-H., Zhou, W.-P., 2014. Antisense long non-coding RNA PCNA-

AS1 promotes tumor growth by regulating proliferating cell nuclear antigen in hepatocellular carcinoma. Cancer Lett. 349, 87–94. doi:10.1016/j.canlet.2014.03.029

Yue, P., Forrest, W.F., Kaminker, J.S., Lohr, S., Zhang, Z., Cavet, G., 2010. Inferring the functional effects of mutation through clusters of mutations in homologous proteins. Hum. Mutat. 31, 264–271. doi:10.1002/humu.21194

Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., Eleftheriadis, M., Sinning, C.R., Schnabe, R.B., Lubos, E., Mennerich, D., Rust, W., Perret, C., Proust, C., Nicaud, V., Loscalzo, J., Hübner, N., Tregouet, D., Münze, T., Ziegler, A., Tiret, L., Blankenberg, S., Cambien, F., 2010. Genetics and beyond - the transcriptome of human monocytes and disease susceptibility. PLoS One 5. doi:10.1371/journal.pone.0010693

Zentner, G.E., Tesar, P.J., Scacheri, P.C., 2011. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res. 21, 1273–1283. doi:10.1101/gr.122382.111

Zhang, A., Zhou, N., Huang, J., Liu, Q., Fukuda, K., Ma, D., Lu, Z., Bai, C., Watabe, K., Mo, Y.-Y., 2013. The human long non-coding RNA-RoR is a p53 repressor in response to DNA damage. Cell Res. 23, 340–50. doi:10.1038/cr.2012.164

Zhang, E., Yin, D., Sun, M., Kong, R., Liu, X., You, L., Han, L., Xia, R., Wang, K., Yang, J., De, W., Shu, Y., Wang, Z., 2014. P53-regulated long non-coding RNA TUG1 affects cell proliferation in human non-small cell lung cancer, partly through epigenetically regulating HOXB7 expression. Cell Death Dis. 5, e1243. doi:10.1038/cddis.2014.201

Zhang, L., Yang, F., Yuan, J. -h., Yuan, S. -x., Zhou, W. -p., Huo, X. -s., Xu, D., Bi, H. -s., Wang, F., Sun, S. -h., 2013. Epigenetic activation of the MiR-200 family contributes to H19-mediated metastasis suppression in hepatocellular carcinoma. Carcinogenesis 34, 577–586. doi:10.1093/carcin/bgs381

Zhang, X., Weissman, S.M., Newburger, P.E., 2014. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. RNA Biol. 11, 1–11. doi:10.4161/rna.28828

Zhang, X., Zhou, Y., Mehta, K.R., Danila, D.C., Scolavino, S., Johnson, S.R., Klibanski, A., 2003. A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. J. Clin. Endocrinol. Metab. 88, 5119–26. doi:10.1210/jc.2003-030222

Zhao, H., Zhang, X., Frazão, J.B., Condino-Neto, A., Newburger, P.E., 2013. HOX antisense lincRNA HOXA-AS2 is an apoptosis repressor in all trans retinoic acid treated NB4 promyelocytic leukemia cells. J. Cell. Biochem. 114, 2375–83. doi:10.1002/jcb.24586

Zhao, J., Liu, Y., Zhang, W., Zhou, Z., Wu, J., Cui, P., Zhang, Y., Huang, G., 2015. Long non-coding RNA Linc00152 is involved in cell cycle arrest, apoptosis, epithelial to mesenchymal transition, cell migration and invasion in gastric cancer. Cell Cycle 4101, 1–12. doi:10.1080/15384101.2015.1078034

Zhao, J., Sun, B.K., Erwin, J. a, Song, J.-J., Lee, J.T., 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322, 750–6. doi:10.1126/science.1163045

Zhao, Y., Luo, H., Chen, X., Xiao, Y., Chen, R., 2014. RNA Mapping 1182, 4939. doi:10.1007/978-1-4939-1062-5

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M., Snyder, M., 2010. Genetic analysis of variation in transcription factor binding in yeast. Nature 464, 1187–1191. doi:10.1038/nature08934

Zhu, Y., Yu, M., Li, Z., Kong, C., Bi, J., Li, J., Gao, Z., Li, Z., 2011. ncRAN, a newly identified long noncoding RNA, enhances human bladder tumor growth, invasion, and survival. Urology 77, 510.e1–5. doi:10.1016/j.urology.2010.09.022

Zou, H., Henzel, W.J., Liu, X., Lutschg, a, Wang, X., 1997. Apaf-1, a human protein homologous to C. elegans CED-4, participates in cytochrome c-dependent activation of caspase-3. Cell 90, 405–13. doi:10.1016/S0092-8674(00)80501-2